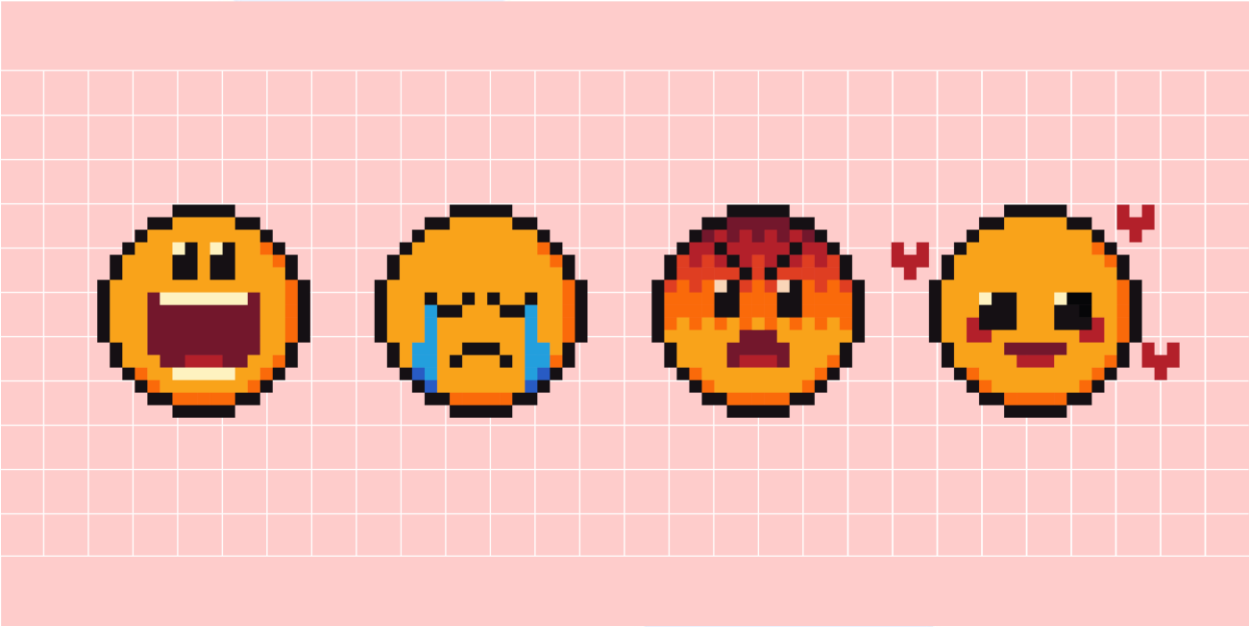
European Commission

Joint Research Centre

Virtual Worlds and Well-being: Current Research and Future Directions

Barreda Ángeles, M., Amendola, S., Da Silva, C., Somià, T., Boucher, P., Schade, S., Meier, A., Mansfield, K., Gunschera, L., Orben, A., Turner, G., Griffiths, M., Hartmann, T., Freeman, G., Gui, X., Kou, Y., Baumgartner, S., Vuorre, M., Ohme, J., Hine, E., López Richart, J., Belmoussi, O.

2025



Contact information
Name: Miguel Barreda Angeles
Address: Via Enrico Fermi 2749 21027 Ispra (VA), Italy
Email: miguel.barreda-angeles@ec.europa.eu
Tel.: +39 03 3278 9437

The Joint Research Centre: EU Science Hub
https://joint-research-centre.ec.europa.eu

JRC143343

PDF ISBN 978-92-68-31434-0 doi:10.2760/7811534 KJ-01-25-462-EN-N

Luxembourg: Publications Office of the European Union, 2025

How to cite this report: Barreda Angeles, M., Amendola, S., Da Silva, C., Somia, T., Boucher, P. et al., *Virtual Worlds and Well-being: Current Research and Future Directions,* Publications Office of the European Union, Luxembourg, 2025, https://data.europa.eu/doi/10.2760/7811534, JRC143343.

Contents

Abstract

Virtual worlds are persistent, immersive environments that blend physical and digital spaces in real time. The expanding use of these technologies underscores the need to study their impact on users' well-being. Drawing on insights from 16 experts across diverse disciplines, this report provides an overview of the current state of scientific knowledge on this topic and outlines pathways for future research and cross-sector collaboration. The expert contributions address a wide range of issues, from lessons learned in previous research on social media and video games, to the distinctive features of virtual worlds and how these may influence users' well-being. They also explore emerging methodological approaches for studying these environments and discuss questions related to platform governance and content moderation. The perspectives gathered in this report highlight the importance of considering technological features, content, and user characteristics to develop a nuanced understanding of the risks and benefits associated with virtual worlds. Integrating these perspectives, this report aims to lay the groundwork for future multidisciplinary research and sustained collaboration among key stakeholders, emphasizing the need for rigorous, inclusive, and policy-relevant studies that can help unlock the opportunities while addressing the challenges of virtual world applications.

Acknowledgments

The Editors would like to express their gratitude to the experts whose active engagement and thoughtful contributions have shaped this volume. We are also grateful to the participants in the seminar series “Virtual Worlds and Well-being: Setting the Research Agenda”, whose discussions and reflections have been instrumental in stimulating debate around the ideas presented by the experts.

Editors

Miguel BARREDA ÁNGELES

Simone AMENDOLA

Cláudia DA SILVA

Tatiana SOMIÀ

Philip BOUCHER

Sven SCHADE

1. Introduction

The European Commission has defined virtual worlds as "persistent, immersive environments, based on technologies including 3D and extended reality (XR), which make it possible to blend physical and digital worlds in realtime, for a variety of purposes such as designing, making simulations, collaborating, learning, socialising, carrying out transactions or providing entertainment" (European Commission, 2023). The earliest forms of digital environments with some of the characteristics of virtual worlds can be found at the very origins of the internet, with examples such as text-based, multi-user dungeons (MUDs) dating back to the 1970s (Downey, 2014). Just a few years later, graphical environments began to emerge (such as the well-known Habitat) and by the 2000s, applications like Second Life or World of Warcraft demonstrated the strong appeal these environments could have for online socialization, entertainment, and education. However, it has been in more recent years, with the growing maturity of technologies such as extended reality (XR), artificial intelligence (AI), 5G/6G networks, and blockchain, that their vast potential across an increasing number of sectors has become evident (Hupont Torres et al., 2023). The European Commission (2023) has recognized that virtual worlds may play a central role in the future of digital connectivity, and emphasized the importance of ensuring their development aligns with European values.

Virtual world applications are gaining presence in many users' lives. Notably, platforms in the gaming domain like Roblox, Minecraft, and Fortnite, are used daily by millions of people, the majority of whom are minors[1] [2]. At the same time, virtual worlds built around immersive VR (e.g., VRChat, Spatial.io) are creating new opportunities for social interaction and remote collaboration, and digital twins are gaining significant traction across a growing number of industrial applications[3]. Given the growing adoption of virtual world applications, the need for research to understand their impact on users' mental health and well-being, particularly in the case of minors and other vulnerable users, has become evident. Supporting such research is among the strategic actions within the *EU Initiative on Web 4.0 and Virtual Worlds* (European Commission, 2023).

1.1. From Digital Media to Virtual Worlds: Advancing Cumulative Knowledge

Concerns about the potential impact of digital technologies on users' well-being are not unique to emerging virtual worlds applications. Whenever a new communication medium —from television to video games to social media— has gained popularity, similar concerns about its potentially harmful effects have surfaced. Typically, the adoption of new technologies (and the emergence of related concerns) has progressed faster than scientists' ability to provide evidence supporting or refuting those concerns, before a new emerging technology captured social attention (Orben, 2020). Within this cycle, there has also been a tendency to start from scratch addressing the impact of each new technological wave, "reinventing the wheel" and overlooking the fact that the mechanisms underpinning the effects of one medium may be shared by others. This pattern has hindered the

[1] https://prioridata.com/data/roblox-users/

[2] https://www.statista.com/statistics/1192573/daily-active-users-global-roblox/

[3] https://www.hkdca.com/wp-content/uploads/2025/02/digital-twin-industry-report-hexagon.pdf

development of a robust and cumulative scientific base that could effectively inform public policy on the impact of media technologies on the population (Orben, 2020). Moreover, research on digital media and well-being —particularly in areas such as social media or video games— has been marked by significant fragmentation and the coexistence of multiple and diverse conceptual and methodological approaches (e.g., King et al., 2013; Meier & Reinecke, 2021). Such heterogeneity has limited the comparability of findings and impeded the cumulative synthesis of evidence.

Against this backdrop, future research efforts should develop a more cumulative, multidisciplinary, and integrative understanding on the impact of virtual worlds on users. This requires taking stock of current relevant evidence across disciplines, identifying the main open questions, and selecting the most appropriate methods to address them. A shared set of priorities can not only help orient future studies but also enable funding bodies to allocate resources more strategically and effectively.

This objective has guided the initial phase of the **VirtueS**[4] project, led by the Joint Research Centre (JRC) of the European Commission. The project's core mission is to advance our understanding of the potential impact of emerging virtual worlds on citizens and European industry from a human-centric perspective. VirtueS is part of the Centre for Advanced Studies programme (within the Scientific Development Programmes unit), which aims to strengthen the JRC's capacity to address the complex and long-term societal challenges facing the European Union.

1.2. About this report

To map existing knowledge and define priorities for future work, the VirtueS project began by engaging in dialogue with experts from a wide range of disciplines —including social psychology, communication science, human–computer interaction, and law— who provided their perspectives as a basis for discussion and further development. This report compiles the outputs of these expert contributions, which also informed the debates held during the online seminar series "Virtual Worlds and Well-being: Setting the Research Agenda", organised by the JRC throughout the first half of 2025. The key insights derived from this process are summarised in the final section of the report.

The expert contributions in this report are organized into four thematic sections. The first section ("**Research on Digital Media and Well-being: Past and Future"**) aims to provide a general overview of the current state of research on the broader topic of digital media and well-being (putting the focus on the types of digital media most scrutinized in the past, like social media and video games), the main lessons learned from the past decades of investigation, and several conceptual and methodological proposals for its future development. Within this section, the report by **Adrian Meier**, titled *Measuring Digital Media in Research on User Mental Health: What Do We Study? And What Should We Care About?*, reviews how digital media use has been examined in relation to mental health, highlighting key conceptual and methodological limitations. It advocates for a communication-centred perspective that considers content, platform experiences, and technology design, offering future directions to better understand the complex interactions between users and digital environments. In turn, the piece *Strategies to Improve Policy Translation from Research on Emerging Technologies*, by **Karen L. Mansfield**, highlights the need for more rigorous evidence on the impact of digital technologies on young people's well-being, and proposes a

[4] https://joint-research-centre.ec.europa.eu/projects-and-activities/centre-advanced-studies/virtues-virtual-worlds-and-society-project_en

structured research framework to identify harms and benefits. It stresses the need for robust methodologies and an open-access resource hub to support evidence-based policymaking. The report by **Lukas J. Gunschera, Amy Orben**, and **Georgia Turner**, *Addressing Gaps in Digital Media Research: Pathways for Future Science and Policy*, identifies some of the main challenges in researching the link between social media use and mental health, and underscores how methodological limitations (like inconsistent measurements and reliance on self-reports) and lack of access to objective data hinder evidence-based policymaking. It proposes moving beyond a "dose-response" model to one that considers the quality and context of digital interactions, and stresses the need for better collaboration between researchers and policymakers. Finally, the piece by **Mark Griffiths**, *Gaming Disorder: A Short Report,* reviews the state of the art on gaming disorder, now recognized as a mental health condition. The author explores how gaming disorder emerges from the interaction of gaming-related, individual, and environmental factors, and proposes innovative paths for advancing future research on this topic.

The second section of this report, "**Unpacking the Impact of Virtual Worlds"**, focuses on those specific features of virtual worlds that are likely to have the greatest impact on user well-being. In this section, the report by **Tilo Hartmann** (*How Presence Shapes the Immersive XR Experience and Potential Wellbeing Effects*) explores the psychology of immersive experiences in virtual worlds accessed through XR technologies, showing how both feelings of presence ("being there") and media awareness jointly determine users' emotional and behavioural responses. The piece *Novel Opportunities and Emerging Risks of Social Virtual Reality Spaces for Online Interactions*, by **Guo Freeman**, in turn, focuses on how social VR platforms such as VRChat and Horizon Worlds are reshaping online social interaction through immersive, avatar-based experiences. Freeman highlights both new opportunities and emerging risks —such as online harassment and shifting power dynamics— and advocates for the design of safer, more inclusive social VR spaces, particularly for marginalized communities. Next, the report by **Xinning Gui** (*Harmful Design Practices in Virtual Worlds*) examines how the democratization of virtual world design has enabled rich user-generated content but also introduced design practices that negatively affect user well-being. The report emphasizes that such harms —ranging from privacy violations to financial loss— often result from complex system interactions rather than individual intentions, and it calls for strengthening ethical agency in virtual world design to address these challenges. **Yubo Kou**'s report, *The Ethical, Governance, and Moderation Aspects of Virtual Worlds*, explores how ethics, governance, and moderation intersect to shape user well-being in these environments. His analysis proposes proactive, value-driven approaches that go beyond enforcement, integrating ethical design, adaptive governance, and community-informed moderation to ensure safety, inclusion, and accountability.

The third section of this volume ("**Towards Methodological Innovation"**) turns to some of the emergent methodological approaches that may contribute to more effective research on emerging virtual worlds. In this section, **Susanne Baumgartner**'s report, *Vanishing Effects? On the Importance of Assessing the Effects of Virtual Worlds on Users' Well-Being Among Early Adopters and First-Time Users*, argues that the impact of digital media on well-being may have been underestimated due to the use of linear media-effect models by researchers. Drawing on psychological theories of habituation and adaptation, Baumgartner defends the need for a new generation of studies focusing on first-time users in order to capture the initial —and potentially strongest— effects before they stabilize over time. **Matti Vuorre**'s report, *Credible and Transparent Industry-Academia Collaborations for Understanding Life Online*, emphasizes that the digital footprints collected by online platforms offer a promising opportunity for studying human behaviour and psychological functioning. This contribution calls for collective efforts to tackle the technical, legal, and ethical challenges currently hindering transparent and responsible collaboration between

academia and industry. Finally, and also related to the use of user data held by private companies, **Jakob Ohme**'s report on *Data Donation as Method: Rethinking Access to Digital Trace Data* presents data donation as a viable approach for studying platform usage and its impact on users, offering an alternative to increasingly restricted Application Programming Interfaces (APIs). It explores how users' data rights can help drive research forward, while analysing the main challenges and identifying possible solutions.

The fourth section (**"Addressing Challenges"**) shifts the focus away from problematizing virtual world use and toward approaches for addressing existing challenges and developing effective solutions. In this section, **Emmie Hine's** report, *From the AI Act to an XR Act? Assessing EU Policy for XR Safety and Privacy*, explores how current EU legal frameworks apply to extended reality (XR) technologies and highlights regulatory uncertainties' resulting from the unique risks and affordances of immersive environments. **Julián López Richart**'s report (*Content Moderation in the Metaverse: Legal Framework and Practical Challenges*) zooms in on content moderation from a legal perspective and the novel challenges that the concept of the metaverse[5] brings to the table. The final contribution, in turn, stands out in tone and approach: **Ouassima Belmoussi**'s report (*Toxicity in Gaming and Virtual Environments: User Perspectives and Needs*) addresses the concept of toxicity in online gaming communities from a dual perspective. On one hand, it presents findings from a study on strategies that gamers apply to cope with toxicity. On the other, it integrates her personal perspective as a gamer, woman, and person of colour, offering a vivid, first-person account that complements and brings to life the topic at hand.

Finally, the section **Key Takeaways and Next Steps** aims to distil the main ideas expressed by the experts in their contributions and debated during the seminar series "Virtual worlds and Wellbeing: Setting the Research Agenda", and to provide an overview of future research activities within the VirtueS project along these lines.

[5] Given the multiple definitions that exist in the academic field for terms such as "virtual worlds" or "metaverse", and the lack of scholarly consensus on their use, the editors of this report decided to respect the terminology as used by the various expert contributors.

2. Research on Digital Media and Well-being: Past and Future

2.1. Measuring Digital Media in Research on User Mental Health: What Do We Study? And What Should We Care About?

Adrian Meier

Friedrich-Alexander-Universität Erlangen-Nürnberg

Abstract

This short report examines how researchers have studied digital media in relation to mental health and uncovers key conceptual and methodological challenges. Five main approaches to measuring *digital media use* (DMU) are highlighted: behavioral, cognitive, clinical, social psychological, and sociological. The popular social psychological active-passive approach to social media use is discussed as an example, as it oversimplifies DMU by failing to account for the specific interactions, contents, and design elements that characterize digital media environments.

The contribution advocates for a shift toward a communication-centered approach to digital media effects that considers *technology design* as a crucial boundary condition for mental health effects. Specifically, the contribution outlines three potential future research foci that might help overcome prior research's common challenges: it advocates for studies on (i) message effects, (ii) cross-platform experiences, and (iii) technology features and their perceived affordances. Together, these approaches understand DMU as a complex interaction between users and technology and centers on the interplay of contents, technology, and social networks in shaping user health.

Finally, the contribution briefly notes key steps ahead for future research on DMU and mental health, such as defining meaningful benchmarks for digital media effects, achieving a balance between specificity and generalizability when conceptualizing DMU, and addressing potential biases like technological determinism and negativity bias in research frameworks. Overall, the contribution aims to provide an understanding of how research has tackled digital media effects on mental health so far – and how it might do so productively in an everchanging future technology landscape.

Highlights

- Identifies five key approaches to studying digital media use and mental health.
- Critiques the active-passive model as case study for an overly simplistic approach.
- Proposes a shift to a communication-centered approach focusing on interactions, content, and design.

Introduction

Researchers and the public controversially debate whether and how ***digital media use*** **(DMU**) – for example, using smartphones, social media, or video games – might impact users' mental health, especially concerning children and adolescents (Kaye et al., 2020; Meier and Reinecke, 2021; Odgers, 2024; Orben et al., 2024; Orben and Blakemore, 2023; Valkenburg, Meier, and Beyens, 2022; Vanden Abeele, Halfmann, and Lee, 2022). Discussions about such "online harms" often revolve around **popular hypotheses**, such as *displacement* of more meaningful and beneficial

activities, e.g. face-to-face interaction or physical activity; *addictive or problematic usage* marked by drastic loss of control over usage; users experiencing *bullying* and other *harmful social interactions*, such as hate speech; unflattering *upward social comparison*, especially regarding body image; impairments of *sleep quality* due to evening media usage; *digital stress* resulting from constant connectivity; and various other potential links between DMU and mental health. Rather than reviewing these manifold individual research lines, this short report examines the **broader underlying patterns of how researchers in this area have studied digital media** in relation to mental health. In doing so, we can identify key conceptual and methodological challenges that help inform future investigations into the social and psychological impacts of emerging digital technologies, such as virtual worlds and AI companions.

First, the short reports distills **five main approaches to measuring *digital media use* (DMU)**: behavioral, cognitive-affective, clinical, social psychological, and sociological. As an example illustrating how each approach comes with unique strengths and weaknesses, the popular **active-passive model of social media use** is discussed in-depth. This model offers an elegant narrative for how social media may both harm and benefit mental health, yet it oversimplifies social media by failing to account for the specific interactions, contents, and design elements that characterize all digital technology environments. My short report thus advocates for a **shift toward an integrated communication-centered approach** to digital media effects that centers on social interaction and message characteristics but considers technology design as a crucial boundary condition for mental health effects. Based on the reviewed research, future scholarship needs to start from the recognition that digital media effects are necessarily complex. They result from an interaction of multiple factors, including user characteristics, specific usage behaviors (e.g., interactions and contents that users engage with), technology design (e.g., certain features and algorithms), and the broader social context (e.g., network and platform structures).

In shifting toward this new approach, the short report points to **key steps ahead for future research on DMU and mental health**, such as achieving a balance between specificity and generalizability when conceptualizing DMU and addressing potential biases like technological determinism and the negativity bias inherent in some current research frameworks. Overall, the contribution aims to provide an understanding of how research has tackled digital media effects on mental health so far – and how it might do so productively in an everchanging future technology landscape.

Five Main Approaches to Studying Effects of DMU on Mental Health

Research on DMU is often concerned with the effects of media technologies on users. Commonly, **media effects** are defined as "the deliberate and non-deliberate short and long-term individual or collective changes in cognitions (including beliefs), emotions, attitudes, and behavior that result from media use" (Valkenburg, Peter, and Walther, 2016, p.316). Hence, (digital) media effects vary along dimensions of *intentionality* (effects intended by users vs. unintended "side effects"), *temporality* (short- vs. long-term), *social organization* (impacts on individuals vs. groups or societies), *psychology* (changes in affect, behavior, cognitions etc.), and *causality*, that is, whether the effects truly "result from media use" or are caused by some confounding third variable (e.g., personality, prior mental health problems, socioeconomic status, etc.)

Existing research has followed various conceptual and methodological approaches to identifying digital media effects on mental health, each emphasizing different elements of this media effects definition and following different disciplinary and epistemological traditions. Based on a cross-cutting synthesis of my prior research (Meier, 2022; Meier, Domahidi, and Günther, 2020; Meier and Reinecke, 2021; Orben et al., 2024; Reinecke et al., 2018; Valkenburg, Meier, and Beyens, 2022), I suggest there are at least **five prototypical approaches to DMU and mental health**:

(1) Behavioral or Technology-Centered Approach: This approach focusses on *observable and quantifiable aspects of DMU* such as whether a device or app is used at all (use vs. non-use), time spent on devices or applications, frequency or regularity of use, and situational patterns of use (e.g., rapid checking, prolonged sessions, fragmented use). Examples can be found particularly in neobehaviorism, epidemiological and public health research, such as investigations into screen time, media multitasking, or media use and sleep (Ahmed et al., 2024; Orben, 2020; Przybylski and Weinstein, 2017; Twenge, 2019; Wiradhany and Koerts, 2021).

(2) Cognitive-Affective or User-Centered Approach: This approach measures *some psychological dimension of DMU*, such as user perceptions, motivations, or attitudes. Thus, this approach centers the user more so than the technology. Examples would be social media mindsets (e.g., believing that social media are harmful or helpful) or motivations for playing video games (Johannes, Vuorre, and Przybylski, 2021; Lee and Hancock, 2024). Sometimes, this approach is combined with the behavioral approach, for example, when studies link observable usage behaviors with users' processing of what they experience in digital environments.

(3) Clinical or Diagnostic Approach: This approach emphasizes *problematic, excessive, or compulsive engagement* with digital media, often framed through the lens of behavioral addiction. Typically, the identification of "addictive use" goes beyond observing high levels of digital media use (e.g., > 8 hours per day) and instead relies on diagnostic scales taken from behavioral addiction (e.g., gambling) or substance abuse (e.g., alcohol) research (Billieux et al., 2015; Fournier et al., 2023; Kardefelt-Winther et al., 2017). These scales are then applied par for par to the digital technology context (e.g., Facebook addiction, smartphone addiction, Internet use disorder).

(4) Social Psychological Approach: This approach examines interpersonal, masspersonal, or mass communicative uses (O'Sullivan and Carr, 2018) of digital media and focusses on *users mediated social interactions with each other and/or with content*, be it from users' social contacts (e.g., friends or family), (semi-)professionally produced (e.g., influencer content), classic storytelling (e.g., viewing TV shows on streaming platforms), or interactive stories (e.g., gaming). One common example examined in more detail below distinguishes between active (e.g., direct messaging) and passive (e.g., browsing content) engagement with social media and their differential effects on mental health (Verduyn et al., 2017).

(5) Sociological Approach: This approach considers *broader systemic and structural aspects of DMU* as embedded within social networks, social groups and societies. Examples include research into the composition of users' entire social media ecosystem, social norms of connection or disconnection (e.g., availability norm, digital stress, phubbing), or media multiplexity, that is, the combination of various technologies people use to stay socially connected (Carter et al., 2023; Reinecke et al., 2018; Taylor, Zhao, and Bazarova, 2022; Triệu et al., 2019).

Depending on the approach, researchers focus on very different facets and measures of DMU, different mechanisms linking DMU to mental health, and different methods for data collection. Each approach exhibits unique strengths and weaknesses. This is illustrated in the following, using the case study of the active-passive model of social media use as an example for the Social Psychological Approach.

Case Study for the Social Psychological Approach: The Active-Passive Model

The *active-passive model* focusses on social media use (SMU) and dichotomizes it into "active" (e.g., engaging with others via private messages or comments) and "passive" (e.g., consuming content without interaction) modes of engagement with the technology (Verduyn et al., 2017). This approach has become popular among researchers, as it allows for a neat reduction of the complexity of SMU into two seemingly distinct modes of engagement. Researchers commonly hypothesize that active SMU should lead to improved well-being and mental health through gains in social resources and connectedness (i.e., *the active use hypothesis*). In contrast, passive SMU is expected to decrease well-being and mental health by eliciting upward social comparison and envy, which deflate users' self-esteem (i.e., *the passive use hypothesis*).

The active-passive model has been highly influential and generated much empirical research. The most comprehensive meta-analysis by Godard and Holtzman (2024) compiled 897 effect sizes (562 active and 334 passive) from 141 studies, including data from over 145,000 participants, mostly from observational and cross-sectional designs. Yet, consistent with prior reviews (Hancock et al., 2022; Valkenburg, van Driel, and Beyens, 2022; Yin et al., 2019), the results from this meta-analysis largely contradict the active-passive model and challenge the utility of the underlying active-passive distinction. For example, both active *and* passive SMU were related to greater perceptions of social support online, which were the strongest effects. Active use showed small positive associations with well-being and positive affect but, contrary to expectations, also with anxiety symptoms. Passive use only showed a small positive association when the analysis pooled all ill-being indicators and excluded social media groups, largely contradicting the passive use hypothesis. Overall, the evidence points to substantially more complex and contingent effect patterns, which has led to an initial revision in the form of an *extended active-passive model* (Verduyn, Gugushvili, and Kross, 2022).

Despite these extensions, the active-passive model still suffers from at least five crucial issues that serve to illustrate the broader challenges for meaningful research into DMU and mental health (Meier et al., 2024), summarized below in Table 1.

Table 1 Five Conceptual Challenges for the Active-Passive Approach

No.	Challenge	Main arguments
#1	"Active" and "passive" do not reflect how social media are actually used	• prominent SM activities (e.g., likes, hearts, shares) cannot be mapped onto active-passive • there is confusion over which activities count as active or passive • active use is better understood as interactive use • passive use is not truly passive, but characterized by selective exposure, engaged viewing, and intentional non-clicks
#2	Active and passive are overly expansive concepts and distract from more nuanced ones	• both active and passive use conflate too many and too different aspects of social media use: types of content, interactions, and platform features • conflating these aspects renders meaningful predictions for effects on mental health impossible • studying these more nuanced aspects (e.g., content, features) promises more robust insights into mental health effects
#3	Active-passive neglects the multiplatform ecology and its evolution	• the active-passive approach does not account for today's multiplatform reality and users' personal social media ecosystems • the active-passive approach is ill-equipped to adapt to an ever-evolving social media landscape
#4	Active-passive invites an incorrect mapping of mechanisms	• passive use can contribute to a sense of social connection • active and passive uses may interact in creating social comparison and envy effects
#5	Active-passive ignores self-effects	• crafting and sending messages can influence the senders themselves, even without reciprocity (i.e., self-effects) • message content and platform features may modulate self-effects, but this is ignored in the active-passive approach

Note. Table 1 was taken from a recent unpublished preprint (Meier et al., 2024).

Source: Author's own elaboration

Broader Conceptual and Methodological Challenges of DMU Research

Zooming out again of this case study, we can identify several broader challenges that future research into DMU and mental health should consider. I highlight four conceptual and methodological challenges, which cut across the main approaches to DMU and mental health outlined above. Each of the main approaches (e.g., behavioral, social psychological, clinical) grapples with at least one of the following broader challenges:

(1) Technological Determinism: This refers to the (often implicit) assumption that technology alone determines user outcomes, such as changes in mental health. Yet, especially when it comes to complex dynamic systems such as mental health, this is a fallacy. Mental health is multicausal and technology is not a monolith. The proposition that a digital technology (e.g., a virtual world or social media platform) produces a uniform, direct, monocausal effect on something as complex and multifaceted as mental health is therefore difficult to defend. Instead, digital media effects are nearly always characterized by complex interactions between user characteristics, usage behaviors, technology design, and various contextual factors (Orben, 2020; Valkenburg, Peter, and Walther, 2016; Valkenburg and Peter, 2013).

(2) Conceptual and Operational Conflation: A common challenge for studies across approaches is that they often fail to distinguish clearly between the factors that give rise to media effects, that is, user characteristics, usage behaviors, technology design, and social contexts, leading to misattribution of effects or overgeneralized conclusions. In research on social media, this issue has been documented through the lens of the *Hierarchical Computer-Mediated Communication Taxonomy* (Meier and Reinecke, 2021). According to this framework, we can organize engagement with digital technologies along six main levels: (1) *Device*, (2) *type of application*, (3) *branded application*, and (4) *features* describe the technologies (aka channels or media) with increasing detail and nuance. Beyond this *channel-centered approach*, researchers may also measure the social interaction behaviors unfolding via channels at the (5) *interaction* or the (6) *message* level, which together form the *communication-centered approach*. A content analysis of nearly 600 studies found that 51% of the measurement instruments used in research on social media and mental health conflate at least two of these levels. For instance, even causally relating a valid indicator of "passive viewing of the Instagram feed" to a subsequent change in mental health tells us nothing about whether the identified effect is driven by a form of interaction (passive viewing), a feature (the feed), or a specific branded app (Instagram). Only through comprehensive research programs that systematically compare these levels (e.g., cross-platform investigations, design experiments) and use nuanced measures (e.g., leveraging trace data at the feature, interaction, and content level) can we identify what, exactly, about technology use may cause mental health problems. In addition, researchers need to consider the heterogeneity on the side of person in front of the screen, as well (Meier et al., 2023; Valkenburg et al., 2021).

(3) Negativity Bias: Both researchers and the public often disproportionately and a priori expect negative outcomes of digital media use while overlooking positive ones. For instance,

upward comparisons on social media may not just elicit envy but also inspiration, yet research on social media comparisons has largely ignored this positive side (Meier et al., 2020; Meier and Johnson, 2022; Valkenburg et al., 2022). This issue is most apparent for the clinical approach (e.g., smartphone addiction), which – by definition – conflates technology engagement (e.g., smartphone use) with negative mental health outcomes (i.e., addictive behavior and related mental health problems due to smartphone use). If a concept necessarily assumes negative effects of a technology on mental health, and this is reflected in measurement (e.g., diagnostic scales including items that attribute mental health problems to a technology), this leaves little room for falsification (Aagaard, 2021; Billieux et al., 2015; Meier, 2022). An approach that is partial to negative technology effects *a priori* will mainly produce findings supporting this assumption.

(4) Standardization of Measures: Current studies use diverse and often inconsistent measures, reducing comparability and replicability. Research in this area is a long way from being a cumulative science. Returning to the example of the active-passive model, one review identified that no two studies relied on the exact same measures of active and passive use (Valkenburg, van Driel, and Beyens, 2022), which is one explanation for the mixed findings on this model. Additionally, self-reports and behavioral measures of DMU (e.g., digital log or trace data) only correlate moderately with each other, raising serious questions of validity (Parry et al., 2021; Parry et al., 2022; Verbeij et al., 2021). However, this finding also points to the fact that technology- and user-centered approaches (see above) try to explain mental health effects through different aspects of technology engagement. Finally, a key challenge for any research into digital technologies is what has been called the *moving target problem*: technologies evolve faster than we can study them (Bayer, Triệu, and Ellison, 2020). For example, the Passive and Active Facebook Use Measure (PAUM), developed eight years ago (Gerson, Plagnol, and Corr, 2017), now suffers from both problems of utility (Facebook is not the most relevant platform among youth anymore) and validity (features now available on Facebook are missing from this measure, while other features have become less important). Together, these challenges lead to the repeated call for *developing standardized, widely accepted self-report measures* of DMU (e.g., Meier and Reinecke, 2021; Trifiro and Gerson, 2019). Additionally, researchers also need to increase the granularity and validity of measures by combining self-reports with *digital trace data*, such as smartphone or browser logs, screenshots/screen-recordings, donations of data download packages under the GDPR, or platform data access under the DSA (Boeschoten et al., 2022; Ohme et al., 2023; van Driel et al., 2022).

Toward an Integrated Communication-Centered Approach

To address (some of) these challenges, this short report closes by advocating for a renewed focus on the most information-rich and temporally stable units of analysis, that is, users' social *interaction* patterns and the types of *messages* (esp. content) they engage with via digital technologies. This *communication-centered approach* (Meier et al., 2024; Meier and Reinecke, 2021) comes with unique challenges of its own (e.g., data access), but is conceptually well positioned to move research on DMU and mental health toward more precise, robust, and replicable insights. Importantly, the approach can draw from decades of insights in communication science, a field centered on interaction and message properties precisely because these characteristics remain relatively stable

compared to the everchanging technology landscape (Walther, 2013; Walther, 2017). However, to also consider the crucial role of technology design, the communication-centered approach needs to be merged with a focus on *technology features and affordances* rather than the more common but rarely informative measurement of device- or application-level characteristics (see the Hierarchical CMC Taxonomy above). Instead, features and affordances have proven a useful lens to identify the specific aspects of a technology that might have implications for mental health, that can be linked to specific psychological processes, and that could be changed through design or regulatory interventions (Evans et al., 2017; Orben et al., 2024). A full explication of this integrated communication-centered approach goes beyond the goals of this short report and can be found in Meier et al. (2024).

Conclusion

To conclude, what do we study when it comes to digital media and mental health, and what should we care about? This contribution distilled five main approaches (i.e., behavioral, cognitive-affective, clinical, social psychological, and sociological) that researchers have applied to tackle the question of whether and how digital media might affect user mental health. Maybe surprisingly, research in this field has produced little conclusive or actionable evidence supporting or refuting digital media effects on mental health, particularly concerning social media and smartphones (Meier and Reinecke, 2021; Odgers and Jensen, 2020; Orben, 2020; Valkenburg, Meier, and Beyens, 2022). I have illustrated both the specific challenges of prominent lines of research (e.g., the active-passive model) and the underlying broader challenges for all five main approaches to DMU and mental health (i.e., technological determinism, conceptual and operational conflation of units of analysis, negativity bias, and lack of standardized measurement). To overcome these challenges in the future, researchers may want to consider a renewed focus on a few key units of analysis – technology features and affordances, social interaction properties, and message characteristics (Meier et al., 2024) – which are most information-rich, relatively temporally stable, and provide a robust foundation for building causal models that link digital technology design to specific psychological mechanisms and subsequent changes in mental health (Orben et al., 2024).

References

Aagaard, J., 'Beyond the Rhetoric of Tech Addiction: Why We Should Be Discussing Tech Habits Instead (and How)', *Phenomenology and the Cognitive Sciences*, Vol. 20, Issue 3, 2021, pp. 559–572.

Ahmed, O., E. Walsh, A. Dawel, K. Alateeq, D.A.E. Oyarce, and N. Cherbuin, 'Social Media Use, Mental Health and Sleep: A Systematic Review with Meta-Analyses', *Journal of Affective Disorders*, 2024.

Bayer, J.B., P. Triệu, and N.B. Ellison, 'Social Media Elements, Ecologies and Effects', *Annual Review of Psychology*, Vol. 71, 2020, pp. 101–1027.

Billieux, J., A. Schimmenti, Y. Khazaal, P. Maurage, and A. Heeren, 'Are We Overpathologizing Everyday Life? A Tenable Blueprint for Behavioral Addiction Research', *Journal of Behavioral Addictions*, Vol. 4, Issue 3, 2015, pp. 119–23.

Boeschoten, L., J. Ausloos, J.E. Möller, T. Araujo, and D.L. Oberski, 'A Framework for Privacy Preserving Digital Trace Data Collection through Data Donation', *Computational Communication Research*, Vol. 4, Issue 2, 2022, pp. 388–423.

Carter, M.C., D.P. Cingel, J.B. Ruiz, and E. Wartella, 'Social Media Use in the Context of the Personal Social Media Ecosystem Framework', *Journal of Communication*, Vol. 73, Issue 1, 2023, pp. 25–37.

van Driel, I.I., A. Giachanou, J.L. Pouwels, L. Boeschoten, I. Beyens, and P.M. Valkenburg, 'Promises and Pitfalls of Social Media Data Donations', *Communication Methods and Measures*, 2022.

Evans, S.K., K.E. Pearce, J. Vitak, and J.W. Treem, 'Explicating Affordances: A Conceptual Framework for Understanding Affordances in Communication Research', *Journal of Computer-Mediated Communication*, Vol. 22, Issue 1, 2017, pp. 35–52.

Fournier, L., A. Schimmenti, A. Musetti, V. Boursier, M. Flayelle, I. Cataldo, V. Starcevic, and J. Billieux, 'Deconstructing the Components Model of Addiction: An Illustration through 'Addictive' Use of Social Media', *Addictive Behaviors*, Vol. 143, 2023.

Gerson, J., A.C. Plagnol, and P.J. Corr, 'Passive and Active Facebook Use Measure (PAUM): Validation and Relationship to the Reinforcement Sensitivity Theory', *Personality and Individual Differences*, Vol. 117, 2017, pp. 81–90.

Godard, R., and S. Holtzman, 'Are Active and Passive Social Media Use Related to Mental Health, Wellbeing, and Social Support Outcomes? A Meta-Analysis of 141 Studies', *Journal of Computer-Mediated Communication* 1, 2024.

Hancock, J.T., S.X. Liu, M. Luo, and H. Mieczkowski, 'Social Media and Psychological Well-Being: A Meta-Analysis', in S. Matz (ed.), *The Psychology of Technology: Social Science Research in the Age of Big Data*, American Psychological Association, Washington, D.C, 2022, pp. 195–238.

Johannes, N., M. Vuorre, and A.K. Przybylski, 'Video Game Play Is Positively Correlated with Well-Being', *Royal Society Open Science*, Vol. 8, Issue 2, 2021, p. 202049.

Kardefelt-Winther, D., A. Heeren, A. Schimmenti, A. van Rooij, P. Maurage, M. Carras, J. Edman, A. Blaszczynski, Y. Khazaal, and J. Billieux, 'How Can We Conceptualize Behavioural Addiction without Pathologizing Common Behaviours?', *Addiction*, Vol. 112, Issue 10, 2017, pp. 1709–1715.

Kaye, L.K., A. Orben, D.A. Ellis, S.C. Hunter, and S. Houghton, 'The Conceptual and Methodological Mayhem of 'Screen Time'', *International Journal of Environmental Research and Public Health*, Vol. 17, Issue 10, 2020, pp. 1–10.

Lee, A.Y., and J.T. Hancock, 'Social Media Mindsets: A New Approach to Understanding Social Media Use and Psychological Well-Being', *Journal of Computer-Mediated Communication*, 2024.

Meier, A., 'Studying Problems, Not Problematic Usage: Do Mobile Checking Habits Increase Procrastination and Decrease Well-Being?', *Mobile Media & Communication*, Vol. 10, Issue 2, 2022, pp. 272–293.

Meier, A., I. Beyens, T. Siebers, J.L. Pouwels, and P.M. Valkenburg, 'Habitual Social Media and Smartphone Use Are Linked to Task Delay for Some, but Not All, Adolescents', *Journal of Computer-Mediated Communication*, Vol. 28, Issue 3, 2023.

Meier, A., E. Domahidi, and E. Günther, 'Computer-Mediated Communication and Mental Health: A Computational Scoping Review of an Interdisciplinary Field', in S.J. Yates and R.E. Rice (eds.), *The Oxford Handbook of Digital Technology and Society*, Oxford University Press, New York, 2020, pp. 79–110.

Meier, A., N. Ellison, L. Reinecke, and P.M. Valkenburg, 'Beyond Active-Passive: Towards the next Stage of Social Media and Mental Health Research', OSF, 2024.

Meier, A., A. Gilbert, S. Börner, and D. Possler, 'Instagram Inspiration: How Upward Comparison on Social Network Sites Can Contribute to Well-Being', *Journal of Communication*, Vol. 70, Issue 5, 2020, pp. 721–743.

Meier, A., and B.K. Johnson, 'Social Comparison and Envy on Social Media: A Critical Review', *Current Opinion in Psychology*, Vol. 45, 2022.

Meier, A., and L. Reinecke, 'Computer-Mediated Communication, Social Media, and Mental Health: A Conceptual and Empirical Meta-Review', *Communication Research*, Vol. 48, Issue 8, 2021, pp. 1182–1209.

Odgers, C.L., 'The Great Rewiring: Is Social Media Really behind an Epidemic of Teenage Mental Illness?', *Nature*, Vol. 628, No. 8006, 2024, pp. 29–30.

Odgers, C.L., and M.R. Jensen, 'Adolescent Mental Health in the Digital Age: Facts, Fears, and Future Directions', *The Journal of Child Psychology and Psychiatry*, Vol. 61, Issue 3, 2020, pp. 336–348.

Ohme, J., T. Araujo, L. Boeschoten, D. Freelon, N. Ram, B.B. Reeves, and T.N. Robinson, 'Digital Trace Data Collection for Social Media Effects Research: APIs, Data Donation, and (Screen) Tracking', *Communication Methods and Measures*, 2023.

Orben, A., 'Teenagers, Screens and Social Media: A Narrative Review of Reviews and Key Studies', *Social Psychiatry and Psychiatric Epidemiology*, Vol. 55, 2020, pp. 407–414.

———, 'The Sisyphean Cycle of Technology Panics', *Perspectives on Psychological Science*, Vol. 15, No. 5, 2020, pp. 1143–1157.

Orben, A., and S.-J. Blakemore, 'How Social Media Affects Teen Mental Health: A Missing Link', *Nature*, Vol. 614, Issue 7948, 2023, pp. 410–412.

Orben, A., A. Meier, T. Dalgleish, and S.-J. Blakemore, 'Mechanisms Linking Social Media Use to Adolescent Mental Health Vulnerability', *Nature Reviews Psychology*, Vol. 3, 2024, pp. 407–423.

O'Sullivan, P.B., and C.T. Carr, 'Masspersonal Communication: A Model Bridging the Mass-Interpersonal Divide', *New Media & Society*, Vol. 20, Issue 3, 2018, pp. 1161–1180.

Parry, D.A., B.I. Davidson, C.J.R. Sewall, J.T. Fisher, H. Mieczkowski, and D.S. Quintana, 'A Systematic Review and Meta-Analysis of Discrepancies between Logged and Self-Reported Digital Media Use', *Nature Human Behaviour*, Vol. 5, 2021, pp. 1535–1547.

Parry, D.A., J.T. Fisher, H. Mieczkowski, C.J.R. Sewall, and B.I. Davidson, 'Social Media and Well-Being: A Methodological Perspective', *Current Opinion in Psychology*, Vol. 45, 2022.

Przybylski, A.K., and N. Weinstein, 'A Large-Scale Test of the Goldilocks Hypothesis: Quantifying the Relations between Digital-Screen Use and the Mental Well-Being of Adolescents', *Psychological Science*, Vol. 28, Issue 2, 2017, pp. 204–215.

Reinecke, L., C. Klimmt, A. Meier, S. Reich, D. Hefner, K. Knop-Huelss, D. Rieger, and P. Vorderer, 'Permanently Online and Permanently Connected: Development and Validation of the Online Vigilance Scale', *PLoS ONE*, Vol. 13, Issue 10, 2018.

Taylor, S.H., P. Zhao, and N.N. Bazarova, 'Social Media and Close Relationships: A Puzzle of Connection and Disconnection', *Current Opinion in Psychology*, Vol. 45, January 1, 2022.

Triệu, P., J.B. Bayer, N.B. Ellison, S. Schoenebeck, and E. Falk, 'Who Likes to Be Reachable? Availability Preferences, Weak Ties, and Bridging Social Capital', *Information, Communication & Society*, Vol. 22, Issue 8, January 1, 2019, pp. 1096–1111.

Trifiro, B.M., and J. Gerson, 'Social Media Usage Patterns: Research Note Regarding the Lack of Universal Validated Measures for Active and Passive Use', *Social Media + Society*, Vol. 5, Issue 2, 2019, pp. 1–4.

Twenge, J.M., 'More Time on Technology, Less Happiness? Associations between Digital-Media Use and Psychological Well-Being', *Current Directions in Psychological Science*, Vol. 28, Issue 4, 2019, pp. 372–379.

Valkenburg, P.M., I. Beyens, J.L. Pouwels, I.I. van Driel, and L. Keijsers, 'Social Media Use and Adolescents' Self-Esteem: Heading for a Person-Specific Media Effects Paradigm', *Journal of Communication*, Vol. 71, Issue 1, 2021, pp. 56–78.

Valkenburg, P.M., I. Beyens, J.L. Pouwels, I.I. van Driel, and L. Keijsers, 'Social Media Browsing and Adolescent Well-Being: Challenging the "Passive Social Media Use Hypothesis"', *Journal of Computer-Mediated Communication*, Vol. 27, Issue 1, 2022.

Valkenburg, P.M., I.I. van Driel, and I. Beyens, 'The Associations of Active and Passive Social Media Use with Well-Being: A Critical Scoping Review', *New Media & Society*, Vol. 24, Issue 2, January 1, 2022, pp. 530–549.

Valkenburg, P.M., A. Meier, and I. Beyens, 'Social Media Use and Its Impact on Adolescent Mental Health: An Umbrella Review of the Evidence', *Current Opinion in Psychology*, Vol. 44, 2022, pp. 58–68.

Valkenburg, P.M., and J. Peter, 'The Differential Susceptibility to Media Effects Model', *Journal of Communication*, Vol. 63, Issue 2, 2013, pp. 221–243.

Valkenburg, P.M., J. Peter, and J.B. Walther, 'Media Effects: Theory and Research', *Annual Review of Psychology*, Vol. 67, 2016, pp. 315–38.

Vanden Abeele, M.M.P., A. Halfmann, and E.W.J. Lee, 'Drug, Demon, or Donut? Theorizing the Relationship between Social Media Use, Digital Well-Being and Digital Disconnection', *Current Opinion in Psychology*, Vol. 45, 2022.

Verbeij, T., J.L. Pouwels, I. Beyens, and P.M. Valkenburg, 'The Accuracy and Validity of Self-Reported Social Media Use Measures among Adolescents', *Computers in Human Behavior Reports*, Vol. 3, 2021.

Verduyn, P., N. Gugushvili, and E. Kross, 'Do Social Networking Sites Influence Well-Being? The Extended Active-Passive Model', *Current Directions in Psychological Science*, Vol. 31, Issue 1, 2022, pp. 62–68.

Verduyn, P., O. Ybarra, M. Résibois, J. Jonides, and E. Kross, 'Do Social Network Sites Enhance or Undermine Subjective Well-Being? A Critical Review', *Social Issues and Policy Review*, Vol. 11, Issue 1, 2017, pp. 274–302.

Walther, J.B., 'Affordances, Effects, and Technology Errors', *Annals of the International Communication Association*, Vol. 36, Issue 1, 2013, pp. 190–193.

———, 'The Merger of Mass and Interpersonal Communication via New Media: Integrating Metaconstructs', *Human Communication Research*, Vol. 43, Issue 4, 2017, pp. 559–572.

Wiradhany, W., and J. Koerts, 'Everyday Functioning-Related Cognitive Correlates of Media Multitasking: A Mini Meta-Analysis', *Media Psychology*, Vol. 24, Issue 2, 2021, pp. 276–303.

Yin, X.-Q., D.A. de Vries, D.A. Gentile, and J.-L. Wang, 'Cultural Background and Measurement of Usage Moderate the Association between Social Networking Sites (SNSs) Usage and Mental Health: A Meta-Analysis', *Social Science Computer Review*, Vol. 37, Issue 5, 2019, pp. 631–648.

2.2. Strategies to Improve Policy Translation from Research on Emerging Technologies

Karen Laura Mansfield

Oxford Internet Institute, University of Oxford

Abstract

Systematic reviews and critical appraisals of research on the impact of technologies on young people's wellbeing have highlighted the heterogeneity of technology effects, and conclude that most published studies are of insufficient rigour to warrant causal interpretation. This has provided policymakers with weak and inconsistent evidence when considering how best to regulate online technologies to protect young people from potential harms. With online platforms increasingly integrating Artificial Intelligence and other emerging technologies, from ranking algorithms to deepfakes, developing a structured and robust framework is critical to dissociating benefits from harms, including how they relate to individual, situational and platform-dependent factors. This short report sets out recommendations to address key challenges to the fast-moving field of digital technology research, building on more rigorous approaches such as mixed methods, generalisable cohort data, causal inference frameworks, and evidence syntheses in living systematic reviews. Promoting these more rigorous approaches and facilitating policy translation would benefit from building an online repository of open access resources, focused on research on emerging technologies and their impact on the wellbeing of people and society.

Highlights

- Research on emerging technology demonstrates inconsistent and heterogeneous effects.
- Heterogeneity reflects not only individual and situational differences, but also study design.
- Facilitating policy translation requires promoting robust methods with stakeholder involvement.
- Best practice includes mixed methods, causal methodology, and living systematic reviews.
- Recommendations can best be promoted via dynamic, open access online resources.

Introduction

A lack of transparent, robust evidence for elucidating the heterogeneous and dynamic health and wellbeing impacts of digital technologies on young people is contributing to the risk of policymakers, practitioners and parents making ungrounded decisions when aiming to safeguard children and adolescents (Mansfield et al., 2025). In November 2024, the Australian government issued a social media ban for under 16s (Online Safety Amendment (Social Media Minimum Age) Bill 2024, 2024). The bill states "providers of certain kinds of social media platforms must take reasonable steps to prevent children who have not reached a minimum age from having accounts", including all platforms whereby "the sole purpose, or a significant purpose, of the service is to enable online

social interaction between 2 or more end-users". The bill goes against the advice of 140 Australian and International experts (Australian Child Rights Taskforce, 2024), who raised multiple concerns, partly based on lack of consideration for young people's rights and fear of isolating them, but also due to a lack of evidence supporting the policy. Besides these concerns, limiting children and adolescents' social media use might be impossible (Houghton et al., 2015), and minors might instead seek social interaction via other unregulated or unsafe means.

Many schools are enacting similar measures, likely with a combination of beneficial and detrimental effects. Schools might assume smartphones to always be a distraction to learning, although many students also use smartphones to look up educational information during or between lessons (OECD, 2024). While there may be advantages to banning smartphones in schools in terms of students' concentration during well-supervised lessons, prohibiting in-school phone use could negatively impact schools with less facilities for self-study and students with less opportunity to do schoolwork at home, exacerbating inequalities in wellbeing and education.

Although social media can be distracting, and online dangers are especially detrimental for vulnerable younger users, a blanket ban based on an age cut-off could have other negative consequences. With regulations specifying age limits, tech firms could interpret this as a free pass to develop and universally integrate age estimation software. Besides the limited accuracy of software that attempts to verify users' age (United States government: National Institute of Standards and Technology, 2024), these algorithms also raise privacy and security concerns (Australian Government: Department of Infrastructure, Transport, Regional Development, Communications, and the Arts, 2023). Furthermore, regulations based purely on users' age means that technology companies are less likely to be motivated to regulate the harmful content that young people become exposed to when they suddenly come of age.

With the rapid integration of Artificial Intelligence (AI) in online platforms, the need to ensure practical and effective technology regulation is becoming increasingly urgent. Facilitating constructive recommendations from academic research can best take a systematic approach, learning from the strengths and weaknesses of past research on existing technologies' effects, and setting out a framework of research priorities, appropriate guidelines and methods for investigating emerging technologies.

Effect heterogeneity reflects study robustness

Informative syntheses of the research on social media's impact on young people include meta-analyses, systematic reviews, scoping reviews, and critical narrative reviews. Many of these reviews have concluded that the effects of social media on children and adolescents' mental health are highly heterogeneous, revealing variability between studies in direction and size of effects (Eirich et al., 2022; Ivie et al., 2020; Liu et al., 2022; Sanders et al., 2024). This implies a worrying amount of ambiguity regarding the potential harms and benefits of social media, making it difficult to determine effective policies for safeguarding young people.

Disentangling the causes of this heterogeneity is a complex task. The inconsistent findings and varying effects could largely reflect (a combination of) the different types of exposures, outcomes, measures and study populations assessed. A few recent reviews have indeed tested heterogeneity to reveal a selection of moderators of the association between technology use and mental health, including Global North versus Global South (Ghai et al., 2023), demographics such as age or gender (Liu et al., 2022; Mougharbel & Goldfield, 2020), and type of technology exposure (Mougharbel & Goldfield, 2020). Besides the varying effects of digital technology between individuals and

situations, an often neglected account of effect heterogeneity is the design and robustness of individual studies. Two meta-analyses published in 2022 demonstrated how effect heterogeneity reflects type of design (e.g. cross-sectional, longitudinal, with or without accounting for baseline mental health) and general study quality, with smaller effects for more robust designs (Eirich et al., 2022; Li et al., 2022). Meanwhile, several reviews have highlighted the abundance of cross-sectional designs in research on technology effects (Eirich et al., 2022; Orben, 2020a; Oswald et al., 2020; Vidal et al., 2020), and many concluded that there is an urgent need for more high quality, causal studies (Berger et al., 2022; Liu et al., 2022; Odgers et al., 2020; Odgers & Jensen, 2020; Oswald et al., 2020; Sanders et al., 2024).

Causal inference requires high-quality science

Causal inference, and translation to policy especially, is dependent on high-quality science, in both experimental and observational settings, at every step in the design, implementation and interpretation of the research. Key considerations include the research question, the choice of outcome and exposure measures, the sample population, identifying and capturing key confounders to account for shared causes, and careful interpretation that acknowledges the limitations of the data and associated methodology. Failing to give proper consideration at any one of these steps can lead to biased findings, lack of causal insight, and misinformed recommendations.

All of this is true for other research areas beyond technology and wellbeing, but the challenges to causal interpretation are amplified with online effects due to the rapidly changing landscape of online technology and artificial intelligence, as well as added pressure from media panics. The onset of new media has historically resulted in concerns, or panics, regarding impacts on young people (Drotner, 1999). Technology panics can be particularly fast-moving cycles, encouraging speed over accuracy in research, while academics are already incentivised to produce outputs quickly in order to progress their careers (Orben, 2020b). Multiple pressures therefore lead to rushed research, non-robust designs, and over-interpretation of available, non-causal findings, increasing the risk of ungrounded translation to policy.

Recommendations – Integrating effective methodologies

Addressing the current shortcomings will involve prioritising quality over quantity and accuracy as well as speed in research outputs, challenging some of the trends in academic research. There are luckily plenty of examples of robust methodology, relevant to different steps in the design and implementation of research. Ensuring that future research investigating the effects of integrated technology is translational could start by integrating some of these effective approaches. Below, and summarised in Table 2, I set out some recommendations for key stakeholders, drawing on effective methods and describing how these can best be used to improve translation to policy.

Mixed methodology with multi-stakeholder involvement

The first step is ensuring the relevance and validity of the research question from the start, addressing issues that policy makers, young people, parents or schools will find most helpful. Researchers are increasingly involving stakeholders and the public in research, such as consulting patient, carer, or youth advisory boards (Moreno et al., 2021), but often this resembles a tick-box exercise to fulfil a requirement by funders, publishers, or ethical review boards. For example, patient and public involvement (PPI) in health research is still not well integrated, although it is most common in mixed-methods mental health research (Lang et al., 2022).

In digital media research, by the time quantitative surveys have been completed, responses collated, and findings reported, technology has moved on. Surveys alone rarely provide adequate time or suitable free text options for respondents to describe their experience to any level of detail. To ensure that research questions are up-to date, meaningful and translational, it is important to consult a diverse, representative sample of all stakeholders expected to be impacted by the research findings, including policy makers, technology companies, practitioners, and the target population. If all public health study designs took a systematic mixed-methods approach, starting with well-planned, structured interviews or focus groups involving all stakeholders, this could increase the relevance, meaningfulness and validity of the research at its foundations.

Causal methodology and clearly defined estimands

Once the research goals are clear, the next requirement for causally interpretable findings is clearly defining the research question and associated estimand. An estimand can be either theoretical or empirical, where the former first defines the constructs and target population for which the research aims to estimate a causal effect, and the latter specifies the observable data that will be used (Lundberg et al., 2021). For example, when designing an intervention, the relevant estimand is the average treatment effect, where the empirical estimand specifies the sample and measures that will be employed. While most peer-reviewed research does outline general research goals, estimands are too often left undefined, with researchers instead focusing on the available data and measures they are analysing (Lundberg et al., 2021).

Although randomised controlled interventions remain the gold standard for causal inference, for many research questions, it is neither practical or ethical to randomise participants to different exposures (e.g. type of online content or social networks). Therefore, health and social scientists often rely on secondary data analyses using observational, non-randomised designs, bring additional challenges to causal inference from multiple potential sources of bias. Many longitudinal observational designs reduce some of this bias by adjusting for baseline measures of the outcome, as well as a selection of stable and time-varying confounders, for example using Random Intercept Cross-Lagged Panel Models (Hamaker et al., 2015). However, important confounders can be neglected if they are not readily available, risking biased effects and ungrounded causal interpretation.

One solution to minimising bias is the Structural Causal Model framework, which employs Directed Acyclic Graphs (DAGs) to explicitly define theoretical causal estimands and to identify potential confounders before identifying suitable data (Hernan & Robins, 2024; Pearl, 2009). DAG development can best be informed by qualitative work with experts and stakeholders. When important confounders are identified but completely missing from available data, the scientifically robust approach would be to first collect the missing measures, for example augmenting existing cohort data, rather than conducting and reporting an analysis without adjusting for those shared causes.

Living systematic reviews

Alongside nurturing high-quality new research, synthesising and evaluating the most important findings from recent research should become a continuous process, helping to identify limitations and gaps in the evidence. The most informative outputs to achieve this goal would be a series of living systematic reviews (Elliott et al., 2014), including a clear workflow and shared code to enable regular updating by multiple teams. Addressing the significant effect heterogeneity expected for different designs, outcomes, populations, groups, situations, and types of online exposure, it is

essential to synthesise studies in a meaningful way and to assess multiple effect moderators, depending on the scope and quality of recent research.

For findings to be causally interpretable and translatable to policy, it can't be stressed enough how important it is to evaluate the robustness of the research, and to incorporate study quality in subsequent syntheses. In other words, high-quality systematic reviews should rigorously assess study quality or risk of bias based on clearly defined, agreed criteria, ideally testing the extent to which such quality accounts for heterogeneity in effects. Poor quality studies should be excluded from final analyses, and studies with lower risk of bias should be ranked to inform policy recommendations (Mansfield et al., 2025). Checklists have already been proposed for assessing quality or risk of bias in non-randomised studies (Higgins et al., 2024; Sterne et al., 2019), and research on emerging technologies would benefit especially from clear guidelines to facilitate objective evaluation of observational studies. Developing a suitable protocol for such living reviews can best be informed by experts in causal inference, risk of bias assessment (especially in non-randomised studies), systematic reviews and meta-analysis.

Open access resources

Finally, ensuring that best practice methodologies are widely adopted requires that clear guidelines and resources promoting high quality research are made openly available to researchers across the globe, inspired by successful examples like the Open Science Framework (Foster & Deardorff, 2017). Many researchers, especially those in low- or middle-income countries, or in universities facing cuts, have less funding, time, and other resources needed to facilitate robust research practices. Open access resources could help to reduce some of these inequalities, supporting alignment on global research priorities and methodologies related to emerging technologies and their impact on people's wellbeing.

Initially, helpful resources might be collections of previously published high-quality systematic reviews and meta-analyses, synthesising a range of research on different types of technology effects. Reviews and primary studies could be scored, drawing on agreed criteria stemming from successful methodologies (Schlussel et al., 2023), and potentially ranked to form an evidence hierarchy of existing research on policy relevant outcomes (Mansfield et al., 2025). Eventually such an evidence hierarchy could be augmented by a series of living systematic reviews such as described above.

Gradually, open resources could present sets of recommendations, and together a framework, for conducting new high-quality research. Recommendations could outline good practice and guidelines for mixed methods designs, causal inference methodology, measurement validity, sample generalisability, DAG development and statistical modelling. These resources could collate and provide links to different types of materials to ensure accessibility, including published papers on methods, accessible blogs, slides, videos, podcasts, well-commented analysis code and step-by-step guides.

As new research on emerging technologies develops, a repository of up-to-date measures (key outcomes, exposures and confounders) can be built, informed by qualitative work with stakeholders from different groups, regions, and populations. Such a repository should include examples of DAGs outlining how they were developed, and eventually associated published studies describing strengths and limitations of the approach.

Data repositories

Online resources for research on technology and wellbeing could also extend to detailed information and links to application procedures for relevant cohort data, highlighting the sample characteristics and available measures, considering strengths and limitations in terms of their scope and external validity (Vazire et al., 2022). New studies could seek to work with stakeholders to identify new exposure measures or missing confounders, which could later be added to existing cohorts, augmenting their causal interpretability and policy value (Mansfield et al., 2025).

Eventually, open data collected with new cohorts, in both observational and experimental settings, can build on these resources, improving on any shortcomings identified by risk of bias assessment in systematic reviews. Research samples often miss the most vulnerable groups, sometimes the most likely to benefit from the research (Mansfield et al., 2023), or don't collect detailed enough information in order to assess heterogeneity in effects between these groups (Ghai et al., 2023). A mixed methods multi-stakeholder approach could also help to engage harder to reach groups and to identify important demographics and other contextual factors for inclusion in the data (Mansfield et al., 2023).

Ideally self-reported data can be integrated with objective data such as from wearables, mobile devices, administrative records, and online platforms. Seeking consent for triangulation of data can reduce the diversity and representativeness of the resulting analysis sample (Mansfield et al., 2020; Morgan et al., 2020), but linking objective data can both add important context and be used to assess the validity and accuracy of self-report measures. There are multiple opportunities for technology firms to collaborate with independent researchers (see Table 2), due to the large-scale, objective data they hold as well as being at the forefront of technology development (Mansfield et al., 2025). However, access to anonymised data from online platforms by independent researchers can be challenging (Breuer et al., 2023), and so until technology firms are better incentivised to work with independent researchers, consenting models might be more promising in the near future, providing sample limitations can be addressed.

Conclusions

For research on the effects of technology on wellbeing to be informative to policy, designs need to be of high-quality, to facilitate causal interpretation. Clear guidelines and online resources are needed to ensure that research on emerging technologies embraces robust methodology, including mixed methods designs, causal inference frameworks, Directed Acyclic Graphs, valid measures, and generalisable samples. Living systematic reviews could be developed to rank the available evidence for policy translation, and to understand the role of study robustness in accounting for heterogeneous findings. Proposals should clearly define how they plan to involve key stakeholders in the research design, promoting the most policy relevant research questions and identifying exposures and contextual factors. Researchers should precisely define theoretical estimands and collect suitable data before conducting analyses when available measures lack contextual information. Online resources should be open access, facilitating global alignment on research priorities and appropriate methods.

Table 2. Recommendations for key stakeholders for each step in the design and translation of research to policy and practice, based on the stakeholder involvement proposed in Mansfield et al., 2025.

Stakeholders: / **Activities:**	**Policymakers**	**Researchers**	**Technology firms**	**Practitioners**	**Young people and families**
Monitor technological developments	Monitor technology developments for significant developments and enable stakeholder workshops	Monitor technological innovation for significant developments and run stakeholder workshops	Announce upcoming technology developments in advance and attend stakeholder workshops	Attend workshops, inform young people about new technology and discuss any concerns	Attend workshops, learn about new technology and discuss any concerns
Define measures, populations and research questions	Fund qualitative research to identify relevant exposures and populations	Collate and validate relevant measures in diverse populations using surveys and focus groups	Support focus groups with families and independent researchers	Recruit diverse participants for focus groups and surveys	Share relevant experience in focus groups and surveys for research questions
Design living systematic reviews that grow with the evidence base	Commission living systematic reviews that critically evaluate bias in samples, measures, and causal interpretation	Carry out and update systematic reviews ranking evidence, measures and impacts	Collaborate with independent researchers to inform research questions	Collaborate with researchers to inform relevant research questions	Collaborate with researchers to inform relevant research questions
Observational data analyses with well-defined causal estimands	Make de-identified administrative data available to researchers via secure platforms	Apply causal methodology with observational data from research, industry and administrative sources	Provide access to de-identified server data and test causal questions with independent researchers	Collaborate with researchers to inform relevant research questions	Collaborate with researchers to inform relevant research questions

Experimental studies assessing exposure-outcome pairs	Fund experimental research to identify instruments for reducing harms and increasing benefits	Use randomized controlled experiments to test instruments for manipulating exposures	Collaborate with researchers to run public-interest experiments within the user base	Collaborate in research design and help to recruit participants for experimental studies	Collaborate in research design or participate in experimental studies
Design, run and assess effectiveness of small- and large-scale interventions	Fund and support integrated interventions within schools, online platforms and other community settings	Design well-controlled trials to assess promising interventions, ensuring co-production	Provide infrastructure for interventions to test promising instruments for reducing harms and increasing benefits	Collaborate in research design and support administration of interventions in schools and community settings	Collaborate in research design or participate in interventions
Create resources, sharing measures, methods and evidence synthesis	Fund and support online resources to align research and recommendations for technology regulation	Collaborate with global experts in the creation of online resources, tools, standards, and findings	Collaborate in the sharing of transparent findings from internal research streams relevant to policy	Support interpretation of findings to community settings and advise on accessibility of resources	Support interpretation of findings for families and young people and advise on accessibility of resources
Develop policy around technology regulation	Develop practical evidence-based policy in discussion with experts in the synthesis and evaluation of research	Advise policymakers, practitioners, and the public, taking care with causal interpretation of the evidence	Discuss and respond effectively to new regulation with transparent information for users	Support the building of digital literacy around new technology and its regulation, and provide feedback	Learn about the new regulation and its motivation, and provide feedback

Source: Author's own elaboration

References

Australian Child Rights Taskforce, 'Re: Proposed social media bans for children under 16-year olds', 9 October 2024, https://www.westernsydney.edu.au/__data/assets/pdf_file/0016/2052160/Open_letter_re_social_media_bans.pdf

Australian Government: Department of Infrastructure, Transport, Regional Development, Communications, and the Arts, 'Government response to the Roadmap for Age Verification', 4 June 2025, https://www.infrastructure.gov.au/sites/default/files/documents/government-response-to-the-roadmap-for-age-verification-august2023.pdf

Berger, M. N., Taba, M., Marino, J. L., Lim, M. S. C., and Skinner, S. R., 'Social media use and health and well-being of lesbian, gay, bisexual, transgender, and queer youth: Systematic review', *Journal of Medical Internet Research,* Vol 24, Issue 9, e38449, 2022, https://doi.org/10.2196/38449

Breuer, J., Kmetty, Z., Haim, M., and Stier, S., 'User-centric approaches for collecting Facebook data in the 'post-API age': Experiences from two studies and recommendations for future research', *Information, Communication & Society,* Vol. 26, Issue 14, 2024, pp. 2649–2668, https://doi.org/10.1080/1369118X.2022.2097015

Drotner, K., 'Dangerous media? Panic discourses and dilemmas of modernity', *Paedagogica Historica,* Vol. 35*, Issue* 3, 1999, pp. 593–619. https://doi.org/10.1080/0030923990350303

Eirich, R., McArthur, B. A., Anhorn, C., McGuinness, C., Christakis, D. A., and Madigan, S., 'Association of screen time with internalizing and externalizing behavior problems in children 12 years or younger', *JAMA Psychiatry,* Vol. 79, Issue 5, 2022, pp. 393–405. https://doi.org/10.1001/jamapsychiatry.2022.0155

Elliott, J. H., Turner, T., Clavisi, O., Thomas, J., Higgins, J. P. T., Mavergames, C., and Gruen, R. L., 'Living systematic reviews: An emerging opportunity to narrow the evidence-practice gap', *PLoS Medicine,* Vol. 11, Issue 2, e1001603, 2014, https://doi.org/10.1371/journal.pmed.1001603

Foster, E. D., and Deardorff, A., 'Open Science Framework (OSF)', *Journal of the Medical Library Association,* Vol. 105, Issue 2, 2017, https://doi.org/10.5195/jmla.2017.88

Ghai, S., Fassi, L., Awadh, F., and Orben, A., 'Lack of sample diversity in research on adolescent depression and social media use: A scoping review and meta-analysis', *Clinical Psychological Science,* Vol. 11, Issue 5, 2023, pp. 759–772. https://doi.org/10.1177/21677026221114859

Hamaker, E. L., Kuiper, R. M., and Grasman, R. P. P. P., 'A critique of the cross-lagged panel model', *Psychological Methods, 20*(1), 2015, pp. 102–116, https://doi.org/10.1037/a0038889

Hernan, M. A., and Robins, J. M., *Causal Inference: What If*. CRC Press, 2024.

Higgins, J. P. T., Morgan, R. L., Rooney, A. A., Taylor, K. W., Thayer, K. A., Silva, R. A., Lemeris, C., Akl, E. A., Bateson, T. F., Berkman, N. D., Glenn, B. S., Hróbjartsson, A., LaKind, J. S., McAleenan, A., Meerpohl, J. J., Nachman, R. M., Obbagy, J. E., O'Connor, A., Radke, E. G., … Sterne, J. A. C., 'A tool to assess risk of bias in non-randomized follow-up studies of exposure effects (ROBINS-E)', *Environment International,* Vol., 186, 108602, 2024, https://doi.org/10.1016/j.envint.2024.108602

Houghton, S., Hunter, S. C., Rosenberg, M., Wood, L., Zadow, C., Martin, K., and Shilton, T., 'Virtually impossible: Limiting Australian children and adolescents daily screen based media use' *BMC Public Health,* Vol., 15, Issue 1, 2015, 5. https://doi.org/10.1186/1471-2458-15-5

Ivie, E. J., Pettitt, A., Moses, L. J., and Allen, N. B., 'A meta-analysis of the association between adolescent social media use and depressive symptoms', *Journal of Affective Disorders,* Vol. 275, 2020, pp. 165–174. https://doi.org/10.1016/j.jad.2020.06.014

Lang, I., King, A., Jenkins, G., Boddy, K., Khan, Z., and Liabo, K., 'How common is patient and public involvement (PPI)? Cross-sectional analysis of frequency of PPI reporting in health research papers and associations with methods, funding sources and other factors', *BMJ Open,* Vol. 12, Issue 5, e063356, 2022, https://doi.org/10.1136/bmjopen-2022-063356

Li, L., Zhang, Q., Zhu, L., Zeng, G., Huang, H., Zhuge, J., Kuang, X., Yang, S., Yang, D., Chen, Z., Gan, Y., Lu, Z., and Wu, C., 'Screen time and depression risk: A meta-analysis of cohort studies', *Frontiers in Psychiatry,* Vol. 13, 1058572, 2022, https://doi.org/10.3389/fpsyt.2022.1058572

Liu, M., Kamper-DeMarco, K. E., Zhang, J., Xiao, J., Dong, D., and Xue, P. 'Time spent on social media and risk of depression in adolescents: A dose–response meta-analysis', *International Journal of Environmental Research and Public Health,* Vol. 19, Issue 9, 5164, 2022, https://doi.org/10.3390/ijerph19095164

Lundberg, I., Johnson, R., and Stewart, B. M., 'What is your estimand? Defining the target quantity connects statistical evidence to theory', *American Sociological Review, Vol. 86*, Issue 3, 2021, pp. 532–565. https://doi.org/10.1177/00031224211004187

Mansfield, K. L., Gallacher, J. E., Mourby, M., and Fazel, M., 'Five models for child and adolescent data linkage in the UK: A review of existing and proposed methods', *Evidence Based Mental Health, Vol. 23*, Issue 1, 2020, pp. 39–44. https://doi.org/10.1136/ebmental-2019-300140

Mansfield, K. L., Ghai, S., Hakman, T., Ballou, N., Vuorre, M., and Przybylski, A. K., 'From social media to artificial intelligence: Improving research on digital harms in youth', *The Lancet Child & Adolescent Health*, S2352464224003328, 2025, https://doi.org/10.1016/S2352-4642(24)00332-8

Mansfield, K. L., Ukoumunne, O. C., Blakemore, S.-J., Montero-Marin, J., Byford, S., Ford, T., and Kuyken, W., 'Missing the context: The challenge of social inequalities to school-based mental health interventions', *JCPP Advances, Vol. 3*, Issue 2, e12165, 2023, https://doi.org/10.1002/jcv2.12165

Moreno, M. A., Jolliff, A., and Kerr, B., 'Youth advisory boards: Perspectives and processes', *Journal of Adolescent Health*, Vol., 69, Issue 2, 2021, pp. 192–194. https://doi.org/10.1016/j.jadohealth.2021.05.001

Morgan, K., Page, N., Brown, R., Long, S., Hewitt, G., Del Pozo-Banos, M., John, A., Murphy, S., and Moore, G., 'Sources of potential bias when combining routine data linkage and a national survey of secondary school-aged children: A record linkage study', *BMC Medical Research Methodology,* Vol. 20, Issue 1, 2020, 178. https://doi.org/10.1186/s12874-020-01064-1

Mougharbel, F., and Goldfield, G. S., 'Psychological correlates of sedentary screen time behaviour among children and adolescents: A narrative review', *Current Obesity Reports*, Vol., 9, Issue 4, 2020, pp. 493–511. https://doi.org/10.1007/s13679-020-00401-1

Odgers, C. L., and Jensen, M. R., 'Annual research review: Adolescent mental health in the digital age: Facts, fears, and future directions', *Journal of Child Psychology and Psychiatry,* Vol.. 61, Issue 3, 2020, pp. 336–348. https://doi.org/10.1111/jcpp.13190

Odgers, C. L., Schueller, S. M., and Ito, M., 'Screen time, social media use, and adolescent development', *Annual Review of Developmental Psychology,* Vol. 2, Issue 1, 2020, pp. 1–18. https://doi.org/10.1146/annurev-devpsych-121318-084815

OECD, 'Students, Digital Devices and Success', *OECD Education Policy Perspectives, No 102*, 2024, https://www.oecd.org/en/publications/students-digital-devices-and-success_9e4c0624-en.html

Parliament of Australia, 'Online Safety Amendment (Social Media Minimum Age) Bill 2024' November 2024, https://www.aph.gov.au/Parliamentary_Business/Bills_Legislation/Bills_Search_Results/Result?bId=r7284

Orben, A., 'Teenagers, screens and social media: A narrative review of reviews and key studies' *Social Psychiatry and Psychiatric Epidemiology, Vol. 55, Issue* 4, 2020a, pp. 407–414. https://doi.org/10.1007/s00127-019-01825-4

Orben, A., 'The Sisyphean cycle of technology panics, *Perspectives on Psychological Science,* Vol. 15, Issue 5, 2020b, pp. 1143–1157. https://doi.org/10.1177/1745691620919372

Oswald, T. K., Rumbold, A. R., Kedzior, S. G. E., and Moore, V. M., 'Psychological impacts of "screen time" and "green time" for children and adolescents: A systematic scoping review', *PLOS ONE,* Vol. 15, Issue 9, e0237725, 2020, https://doi.org/10.1371/journal.pone.0237725

Pearl, J., 'Causal inference in statistics: An overview', *Statistics Surveys,* Vol. 3, 2009, pp 96–146. https://doi.org/10.1214/09-SS057

Sanders, T., Noetel, M., Parker, P., Del Pozo Cruz, B., Biddle, S., Ronto, R., Hulteen, R., Parker, R., Thomas, G., De Cocker, K., Salmon, J., Hesketh, K., Weeks, N., Arnott, H., Devine, E., Vasconcellos, R., Pagano, R., Sherson, J., Conigrave, J., and Lonsdale, C., 'An umbrella review of the benefits and risks associated with youths' interactions with electronic screens', *Nature Human Behaviour, Vol. 8*, Issue 1, 2024, pp. 82–99. https://doi.org/10.1038/s41562-023-01712-8

Schlussel, M. M., Sharp, M. K., De Beyer, J. A., Kirtley, S., Logullo, P., Dhiman, P., MacCarthy, A., Koroleva, A., Speich, B., Bullock, G. S., Moher, D., & Collins, G. S., 'Reporting guidelines used varying methodology to develop recommendations', *Journal of Clinical Epidemiology,* Vol. 159, 2023, pp. 246–256. https://doi.org/10.1016/j.jclinepi.2023.03.018

Sterne, J. A., Hernán, M. A., McAleenan, A., Reeves, B. C., & Higgins, J. P., 'Assessing risk of bias in a non-randomized study'. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, and V. A. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions* (1st ed., pp. 621–641), 2019, Wiley. https://doi.org/10.1002/9781119536604.ch25

United States government: National Institute of Standards and Technology. *'NIST reports first results from age estimation software evaluation'*. 30 May 2024, updated 4 February 2025, https://www.nist.gov/news-events/news/2024/05/nist-reports-first-results-age-estimation-software-evaluation

Vazire, S., Schiavone, S. R., and Bottesini, J. G., 'Credibility beyond replicability: Improving the Four Validities in psychological science', *Current Directions in Psychological Science, Vol. 31*, Issue 2, 2022, pp. 162–168. https://doi.org/10.1177/09637214211067779

Vidal, C., Lhaksampa, T., Miller, L., and Platt, R., 'Social media use and depression in adolescents: A scoping review', *International Review of Psychiatry,* Vol. 32, Issue 3, 2020, pp. 235–253, https://doi.org/10.1080/09540261.2020.1720623

2.3. Addressing Gaps in Digital Media Research: Pathways for Future Science and Policy

Lukas J. Gunschera, Amy Orben, Georgia Turner[6]

MRC Cognition and Brain Sciences Unit, University of Cambridge

Abstract

This short report examines the challenges in researching the relationship between social media use and mental health, highlighting the methodological and structural limitations that hinder evidence-based policymaking. The current evidence is marked by inconsistent findings resulting from inconsistent measurement approaches, reliance on subjective self-report, and inadequate theoretical frameworks. Three key challenges include the lack of causal evidence, restricted access to objective platform data, and the mismatch between research timelines and technological development. As a result, policymakers and researchers face significant obstacles in anticipating and mitigating the effects of new digital technologies on individuals and society. The dominant "dose-response" framework used to understand technology impacts, which links time spent on social media with outcomes of interest, is often inadequate and does not capture the nuanced and multidimensional nature of digital interactions. Emerging paradigms such as the “digital diet” model emphasise the quality and context of online interactions rather than sheer quantity. Individual differences, platform-specific features, and user experiences must also be considered. The following areas should be targeted to confront these issues: 1) moving beyond simplistic dose-response models to explore the active ingredients of digital environments, 2) addressing the limitations of subjective self-reports and securing better access to objective platform data, and 3) accelerating funding and regulatory processes to keep pace with rapidly evolving digital technologies. Addressing these challenges will require reimaging research priorities, close collaboration between researchers and policymakers, and adapting regulatory frameworks. Only then can policymakers effectively navigate the implications of an increasingly digital world.

Highlights

- Research on the impacts of social media has produced inconsistent and inconclusive results.
- Dose-response frameworks of social media effects are inadequate.
- We need to consider digital platforms’ “active ingredients” to understand their nuanced effects.
- Reliance on self-report and lack of access to objective data hinders research.
- Funding priorities need to be reimagined to facilitate research on developing technologies.

---

[6] Alphabetical order by last name

Brief Overview

The relationship between social media use and adolescent mental health is a complex and contentious area of research. Current evidence is marked by inconsistent findings. These discrepancies are driven by varied approaches to measuring social media use, reliance on subjective self-reports, and the absence of robust theoretical frameworks. Researchers often focus on metrics like time spent online, but such simplistic measures fail to capture the nuanced and multidimensional impacts of digital environments.

Key challenges include the lack of causal evidence and limited access to objective data from social media platforms, which hampers the development of actionable insights. Current research practices are also unable to keep pace with the rapid evolution of digital platforms, making findings outdated by the time they are published. Policymakers and researchers face significant obstacles in anticipating and mitigating the effects of new digital technologies on individuals and society.

The dominant "dose-response" framework used to understand technology impacts, which links time spent on a digital environment with outcomes of interest, is often inadequate. Emerging paradigms, such as the "digital diet" model, emphasise the need to consider the quality and context of online interactions rather than sheer quantity. Individual differences, platform-specific features, and user experiences must also be considered.

To learn from the issues confronting social media research and policymaking, research and policy concerning new and immersive digital environments should focus on the following:

1. Moving beyond simplistic dose-response models to explore the "active ingredients" of digital environments, accounting for individual differences in the process.

2. Addressing the limitations of subjective self-reports by securing better access to objective platform data for more accurate and meaningful research.

3. Accelerating funding and regulatory processes to keep pace with rapidly evolving digital technologies and ensure timely, actionable insights.

Looking forward, addressing these challenges will require reimagining research priorities, fostering closer collaboration between researchers and policy, and adapting regulatory frameworks to the fast-changing digital landscape. Only then can policymakers effectively navigate the implications of an increasingly digital and immersive world.

The current state of evidence on social media and mental health

The relationship between social media and mental health in adolescents has emerged as one of the most pressing policy questions of recent years, commanding scrutiny from researchers, policymakers, and the public alike. While numerous studies have attempted to provide clarity about the impact of this type of immersive digital environment, the field has been characterised by heterogeneous findings (Orben, 2020; Sanders et al., 2023; Tang et al., 2021), including positive, negative or null effects of social media use on mental health.

One reason behind such mixed and inconclusive findings is the heterogenous ways in which researchers conceptualise and measure social media use (Kaye et al., 2020; Sanders et al., 2023). Some measure time spent on a specific platform or the smartphone, others ask for self-reports of ‘addiction’ or problematic use, while a different team might focus on reports of activities or content

engaged with. The lack both of consistent measurement, and of theoretical foundations for what to measure and why, has hampered progress and confidence in resulting conclusions (Orben, 2020).

Further, there is widespread use of subjective (i.e., reported rather than directly tracked) social media measures and assessments. Such measures, for example asking individuals to estimate the time they spend on digital media, are unreliable as people are poor at remembering the time they spend interacting with platforms (Parry et al., 2021).

Moreover, pervasive reliance on correlational research in studying social media's impact significantly limits our ability to discern its causal effects. Without carefully designed experimental or quasi-experimental studies that can isolate and manipulate specific variables, researchers cannot confidently determine whether observed associations are due to social media's direct effects, pre-existing individual differences, or complex bidirectional interactions. For instance, does social media use affect wellbeing or does wellbeing affect the ways we engage on social media? Answers to this and similar questions are vital for developing robust evidence-based policies. As digital platforms continue to develop at an unprecedented pace, policymakers and researchers must collaborate to ensure that future studies employ methodologies that can identify causal mechanisms, thereby providing a foundation for interventions and regulatory frameworks.

To summarise, there are many limitations of research on social media, from lacklustre measurement to an absence of high-quality causal evidence. Further, there are few useful theoretical frameworks that facilitate more meaningful investigation into the nuanced effects of digital environments like social media and enable us to make predictions about how changing digital landscapes might impact population mental health and society in the future. Current research conclusions on such topics are often therefore inconsistent, untimely and of low quality. Policymakers attempting to build on such research struggle, failing to consider the complexity of social media and the online environment it creates. This not only limits our current understanding but also hinders our ability to anticipate and prepare for the implications of new technological developments and innovations in an accelerating digital age.

Time spent and the dose-response relationship

A dominant framework for thinking about the individual effects of digital technologies or environments the dose-response model. This model assumes a direct relationship between a specific ‘dose’ of the digital environment in question and subsequent effects on mental or physical health. Often, the ‘dose’ is conceptualised as a specific amount of time spent using a specific digital environment. The dose-response model is seen in many forms of policymaking and research.

For example, in 2018 UK policymakers considered implementing screen use guidelines for children inspired by those for alcohol consumption (with adults recommended to not exceed a certain amount of alcohol units per week). One would have a recommended amount of time any child should be allowed to spend on screens that is labelled acceptable. This assumes a dose-response relationship, with ‘time spent’ as the active ingredient that is meant to be regulated to alleviate certain negative effects of screen time. However, a Chief Medical Officer report commissioned on this question (Davies et al., 2019) found that this assumption was mistaken (Hawkes, 2019). While alcohol is a molecule with a specific impact on the body for every additional millilitre consumed, screen time is not the same.

The dose-response assumption is also reflected in research examining predominantly the effects of time spent on social media and in parents concerned mostly about the time their children spend on social media (Coyne et al., 2020; Verbeij et al., 2021). This reductionist view originates from medical models that postulate a direct relationship between the quantity of exposure to a substance and its consequent biological or psychological effect (Calabrese, 2016). Transposing this approach onto social media implies the presence of a relationship between the time individuals spend on social media and their mental and physical well-being. However, studies looking at time spent online have failed to provide meaningful insight (Hawkes, 2019).

One reason for this is that digital environments like social media are fundamentally different from pharmacological interventions, rendering simplistic models focused on a dose-response reductionistic and problematic. Instead, thinking about online social environments with the same nuance we attribute to offline social environments stands to be a more productive and meaningful framework for considering their potential positive and negative effects (Orben et al., 2024). The metaphor of 'digital diet' might provide such a framework, suggesting that digital engagement is less about quantity and more about the composition, quality, and individual metabolisation of digital experiences (Orben, 2022). The perspective challenges the dose-response model by highlighting the complexity and multidimensionality of online social experiences, emphasising that the same amount of time spent on social media can have dramatically different impacts across different individuals, platforms, and contexts.

While acknowledging the complexity of digital environments, it is nevertheless crucial to recognise that some aspects of digital environments may exhibit dose-response relationships (Kowalski et al., 2014). For example, exposure to harmful content, such as pro-anorexia material, has been linked with negative outcomes like disordered eating behaviours, with effects that intensify with greater exposure (Fardouly & Vartanian, 2016). Moreover, some research has demonstrated dose-response patterns between social media use and relevant outcomes, such as risk-taking behaviours (Purba et al., 2024).

If we collectively decide that the dose-response model cannot be the sole approach to research and policy for questions on digital environments, this requires a fundamental reimagining of the ways in which we conceptualise both research and policy. The limitations of the dose-response model become evident when we consider the multifaceted and dynamic nature of digital interactions, the causal structure of which cannot adequately be captured using oversimplistic frameworks. Aggregate-level inquiries, such as "what is the effect of social media on mental health", should not be the only questions of interest as they fail to appreciate the complexity of online social environments and stand to yield little meaningful insight if not approached with care. Instead, careful experimental and intervention research designs capable of drawing causal conclusions need to be prioritised. Doing so will facilitate a nuanced approach to digital media research and policy that enables meaningful insight needed to examine and predict the effects of specific digital design features or experiences.

Such a shift requires a parallel transformation in both research and funding ecosystems. Funding bodies and governments must develop a more nuanced approach to funding and understanding research that goes beyond expecting straightforward, generalisable answers to very broad questions. Instead, research that is both practically relevant and scientifically rigorous should be encouraged. Further, researchers need to avoid overgeneralisation and better communicate the intricate interactions between digital platforms, individual characteristics, and psychological outcomes. This involves supporting and developing innovative methodological approaches and

facilitating interdisciplinary collaboration, which can in turn be supported by an adaptive funding ecosystem.

Objective social media data

One of the most substantial barriers to providing high quality evidence moving beyond time spent on social media is the reliance on self-report measures (Orben & Przybylski, 2019; Twenge & Campbell, 2018). Retrospective self-reports of media use correlate only moderately with objective usage levels, measured for example through screen time logs (Parry et al., 2021). Importantly, social media use measured through self-report has a stronger correlation with self-reported wellbeing outcomes than objectively measured use (Parry et al., 2021). Thus, research using self-report measures cannot adequately tease apart the effects of actual vs. perceived media use on wellbeing outcomes.

A potential remedy to this reliance on self-report measures lies in the vast amounts of objective digital footprint data social media users create every day - for example, records of posts made, direct messages sent, or videos viewed. Social media companies use such objective data for internal research in both observational and intervention studies. Recently, companies have also started using such data to publish external research. For example, in the '2020 Election project', internal researchers within Meta collaborated with external academic researchers in a field experiment which manipulated Facebook algorithms in order to study their effects on political polarisation (Wagner, 2023).

In contrast, company-external researchers such as academics can often only access objective social media data via platform-controlled Application Programming Interfaces (APIs), if at all. This system of data access limits the scope and reproducibility of findings. Such APIs are subject to unexpected changes, such as the shutting down of CrowdTangle, Meta's public insights tool, in 2024, and the introduction of prohibitive financial costs for use of the API of X (formerly Twitter) when ownership changed in 2022. Given that companies are free to alter their policies at any time to limit new data access, previous findings based on such data can be rendered impossible to reproduce. Moreover, poor communication of such policy changes can create ambiguity and confusion about what is permitted. Finally, even when such research is carried out, restrictive platform policies on how the data can be used, including preclusion of sharing and mandatory deletion after certain time periods, restrict the ability of researchers to openly share their work (Davidson et al., 2023). Obtaining research data via processes other than platform-controlled APIs, such as web-scraping, risks legal consequences, as when Meta took legal action against researchers who developed the Ad Observer research tool in 2020 (Kenny, 2021).

In this context, a key opportunity for policy to advance social media research lies in facilitating external researcher access to, and opportunities to perform experimental interventions on, objective social media data. Regarding data access, novel policy could mandate researcher access to objective data not only on social media users' activities, but also on internal processes such as the results of internal research, as well as algorithmic and content moderation processes. The recently released draft of the European Union Digital Services Act shows promising plans in all these areas, as well as a centralised data access portal, which could improve transparency and consistency of data access. While access for non-EU researchers will necessarily be limited by GDPR, we recommend that regulation works towards equitable access for all external researchers regardless of location, to allow the international collaboration necessary to produce high quality research.

However, the gold standard for causal inference is the practice of randomised experimental interventions. This is exemplified in internal company research using A/B testing of design features,

where social media users are randomly allocated to different social media environments, and the differential effects of these separate conditions are measured to select the option that maximises the outcome of interest (e.g., time spent on platform). In the current landscape, internal researchers within social media companies perform all design and implementation of social media platform product testing. This contrasts, for example, with vehicle safety testing, where external, government researchers have the power to design and conduct product tests before vehicles are designated appropriate for the market (implemented via independent governmental bodies such as the Vehicle Certification Agency in the United Kingdom, as mandated by the United Nations Economic Commission for Europe 1958 Agreement). Importing this model to the social media research domain could provide unprecedented opportunities for large-scale experimental research on social media data, helping to address the gap in causal evidence about the effects of specific aspects of social media, and other digital environments in future, on outcomes such as mental wellbeing.

Research speed and its issues

While access to data and experimental resources could greatly increase the quality of evidence, research must also produce results at a fast enough pace to inform policy. It is particularly difficult to provide sufficiently fast evidence in research on emerging technologies, given the constantly evolving objects of study (Orben, 2020). Scientific studies typically take at least 2-3 years from ideation, funding acquisition, implementation to publication. This timescale is a mismatch with the rate of change of social media platforms, which fluctuate in popularity and design at often quicker and unpredictable paces. Scientific findings can therefore become out of date even by the time they are published.

Policy could begin to address this timescale mismatch in at least three actionable ways. First, research funding priorities could alter their criteria to better accommodate the rapidly changing nature of the digital world. Conventional project-based funding of science requires a problem to emerge, subsequent funding applications for projects on this problem to be written, peer reviewed and funded, and then research teams to be built. This process is inherently conservative and reactive, and therefore ill-equipped for providing timely insights into social media impact. An alternative funding model might prioritise research into the anticipation of changes in digitalisation as or before they happen. For example, policymakers might fund research into how large language models or deepfakes could alter the social media experience and affect young people's development and wellbeing, even if such phenomena are not yet widespread. This prospective funding approach could represent a higher risk at the level of each individual study, but at the cumulative level is more likely to produce the forward-looking evidence needed to confront technological change.

A second policy action could take the form of a rethink about how much evidence can feasibly be provided through the current research and funding system due to the problems above. It might be that very high standards of evidence cannot be reached in time and should not be expected through the system. Thus, it might be necessary to change the requirement for such traditional research in the social media policy landscape (Orben & Matias, 2025).

Finally, policymakers could mandate integrating 'safety by design' approaches into online platforms. Currently, there is minimal regulation to constrain the designs of social media platforms, so that problems are often only researched and addressed after the products have been released to users. In contrast, safety by design means designing platforms with users' safety as a primary objective. This includes measures which prevent harms before they occur, such as high default privacy

settings, as well as limiting access to harmful content, and making reporting processes clearer. Crucially, this safety objective may conflict at times with other design objectives such as maximising profit via increasing users' time spent online. In such cases, policy regulation will be needed to ensure that companies prioritise safety even if this comes at a financial loss. Given the gaps in scientific evidence about social media impact discussed above, safety by design could best be implemented via a close dialogue between companies and researchers. This would allow safety by design principles to be informed by the most current research, and also feed back into research by data access.

Looking forward

What can the current debate and substantial amount of academic research on the impact of social media on adolescent wellbeing and mental health tell us about how we might better do research to inform policy about immersive digital and virtual worlds in the future? In this short report, we argue that there are three important lessons to be learnt.

First, it is often a default to first study the impact of new digital environments through a dose-response lens: examining how much time spent in the environment impacts the outcome of interest. We now know from many previous technologies that this endeavour often fails. Research needs to consider the various 'active ingredients' of the digital environment and how they might in turn impact outcomes, and how individual differences will moderate this relationship. Some of these might be a dose-response relationship, while others will not be.

Second, to do such work well, researchers need to have good measures of the immersive digital environment at their disposal. Currently, a lack of data access agreements with technology companies often leaves researchers no choice but to use questionnaire subjective self-report measures and a focus on time spent to understand their technology of interest.

Third, even if we solve data access issues and provide new theoretical foundations for research, the science-policy ecosystem is facing an accelerating digital world. The current system is outpaced by technological change, due to the slow pace of funding and production of science. Funding, evidence evaluation, and regulatory systems need to urgently adapt to face up to this challenge.

References

Calabrese, E. J.. 'The emergence of the dose–response concept in Biology and Medicine', *International Journal of Molecular Sciences*, Vol. 17, Issue 12, Article 12, 2016, https://doi.org/10.3390/ijms17122034

Coyne, S. M., Rogers, A. A., Zurcher, J. D., Stockdale, L., and Booth, M., 'Does time spent using social media impact mental health?: An eight year longitudinal study', *Computers in Human Behavior*, Vol. 104, 106160, 2020, https://doi.org/10.1016/j.chb.2019.106160

Davidson, B. I., Wischerath, D., Racek, D., Parry, D. A., Godwin, E., Hinds, J., van der Linden, D., Roscoe, J. F., Ayravainen, L., and Cork, A. G., 'Platform-controlled social media APIs threaten open science', *Nature Human Behaviour*, Vol. 7, Issue 12, 2023, pp. 2054–2057, https://doi.org/10.1038/s41562-023-01750-2

Davies, S., Atherton, F., Calderwood, C., and McBridge, M., *United Kingdom Chief Medical Officers' commentary on 'Screen-based activities and children and young people's mental health and psychosocial wellbeing: A systematic map of reviews'*. Department of Health and Social Care, 2019.

Fardouly, J., and Vartanian, L. R., 'Social media and body image concerns: Current research and future directions', *Current Opinion in Psychology*, Vol. 9, 2016, pp. 1–5. https://doi.org/10.1016/j.copsyc.2015.09.005

Hawkes, N., 'CMO report is unable to shed light on impact of screen time and social media on children's health', *BMJ*, Vol. 364, l643, 2019, https://doi.org/10.1136/bmj.l643

Kaye, L. K., Orben, A., Ellis, D. A., Hunter, S. C., & Houghton, S., 'The conceptual and methodological mayhem of "screen time"', *International Journal of Environmental Research and Public Health*, Vol. 17, Issue 10, 2020, https://doi.org/10.3390/ijerph17103661

Kenny, L., *Researchers, NYU, Knight Institute Condemn Facebook's Effort to Squelch Independent Research about Misinformation*, 2021, http://knightcolumbia.org/content/researchers-nyu-knight-institute-condemn-facebooks-effort-to-squelch-independent-research-about-misinformation

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., and Lattanner, M. R., 'Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth', *Psychological Bulletin*, Vol. 140, Issue 4, 2014, pp. 1073–1137, https://doi.org/10.1037/a0035618

Orben, A., 'Teenagers, screens and social media: A narrative review of reviews and key studies', *Social Psychiatry and Psychiatric Epidemiology*, Vol. 55, Issue 4, 2020, pp 407–414. https://doi.org/10.1007/s00127-019-01825-4

Orben, A., 'Digital diet: A 21st century approach to understanding digital technologies and development', *Infant and Child Development*, Vol. 31, Issue 1, 2022, https://doi.org/10.1002/icd.2228

Orben, A., and Matias, J. N.. 'Fixing the science of digital technology harms', *Science*, Vol. *388*, issue 6743, 2025, pp. 152–155, https://doi.org/10.1126/science.adt6807

Orben, A., Meier, A., Dalgleish, T., and Blakemore, S.-J., 'Mechanisms linking social media use to adolescent mental health vulnerability', *Nature Reviews Psychology*, 2024, 1–17. https://doi.org/10.1038/s44159-024-00307-y

Orben, A. and Przybylski, A. K., 'The association between adolescent well-being and digital technology use', *Nature Human Behaviour*, Vol. 3, Issue 2, Article 2, 2019, https://doi.org/10.1038/s41562-018-0506-1

Parry, D. A., Davidson, B. I., Sewall, C. J. R., Fisher, J. T., Mieczkowski, H., and Quintana, D. S., 'A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use', *Nature Human Behaviour*, Vol. 5, Issue 11, 2021, pp. 1535–1547. https://doi.org/10.1038/s41562-021-01117-5

Purba, A. K., Henderson, M., Baxter, A., Pearce, A., and Katikireddi, S. V., 'The relationship between time spent on social media and adolescent cigarette, e-cigarette, and dual use: A longitudinal analysis of the UK Millennium Cohort Study', *Nicotine & Tobacco Research*, ntae057, 2024, https://doi.org/10.1093/ntr/ntae057

Sanders, T., Noetel, M., Parker, P., Del Pozo Cruz, B., Biddle, S., Ronto, R., Hulteen, R., Parker, R., Thomas, G., De Cocker, K., Salmon, J., Hesketh, K., Weeks, N., Arnott, H., Devine, E., Vasconcellos, R., Pagano, R., Sherson, J., Conigrave, J., and Lonsdale, C., 'An umbrella review of the benefits and risks associated with youths' interactions with electronic screens', *Nature Human Behaviour*, 2023, pp. 1–18. https://doi.org/10.1038/s41562-023-01712-8

Tang, S., Werner-Seidler, A., Torok, M., Mackinnon, A. J., and Christensen, H., 'The relationship between screen time and mental health in young people: A systematic review of longitudinal studies', *Clinical Psychology Review*, Vol., 86, 102021, 2021, https://doi.org/10.1016/j.cpr.2021.102021

Twenge, J. M., and Campbell, W. K., 'Associations between screen time and lower psychological well-being among children and adolescents: Evidence from a population-based study', *Preventive Medicine Reports*, Vol. 12, 2018, pp. 271–283, https://doi.org/10.1016/j.pmedr.2018.10.003

Verbeij, T., Pouwels, J. L., Beyens, I., and Valkenburg, P. M., 'The accuracy and validity of self-reported social media use measures among adolescents', *Computers in Human Behavior Reports*, Vol. 3, 100090, 2021, https://doi.org/10.1016/j.chbr.2021.100090

Wagner, M. W., 'Independence by permission', *Science*, Vol. 381, Issue 6656, 2023, pp. 388–391, https://doi.org/10.1126/science.adi2430

2.4. Gaming Disorder: A Short Report

Mark D. Griffiths

International Gaming Research Unit, Nottingham Trent University

Abstract

Gaming disorder (GD) has now been officially recognised as a mental health disorder. GD is the result of an interplay between gaming-related factors, individual factors, and environmental factors. None of these alone are sufficient to cause a disordered state, but it is the interactive co-occurrence of all these factors, which in some cases leads to GD. This short report discusses these three factors based on relevant and recent findings of the literature, and provides recommendations for future research.

Highlights

- A small minority of individuals worldwide appear to develop gaming disorder (GD).
- GD develops due to gaming-related factors, individual factors, and environmental factors.
- Males are much more likely to experience GD than females.
- Comorbidity tends to be the norm rather than the exception (e.g., anxiety, depression) in GD.
- Structural characteristics of the videogames themselves may also contribute to GD.

Background

Research examining problematic videogame playing dates back to the early 1980s when the first reports started appearing concerning adolescents being 'obsessed' with or 'addicted' to the playing of arcade videogames such as Space Invaders (Griffiths et al., 2012). The playing of videogames (i.e., 'gaming') has evolved during this time from playing videogames in amusement arcades in the 1980s, to playing videogames on dedicated gaming consoles and personal computers in the 1990s, to playing videogames online in the 2000s (Griffiths et al., 2012).

More recently, technology has advanced so that gaming can be engaged in from almost anywhere through smartphones and Wi-Fi-enabled mobile handheld devices, as well as in virtual reality (Lopez-Fernandez et al., 2018). Historically, gaming has traditionally been an activity predominantly engaged in by children and adolescents, but gaming has now become a popular activity among adults (Griffiths et al., 2012). However, children and adolescents to be a vulnerable group when it comes to experiencing the negative consequences of gaming excessively and can adversely affect their educational performance, mental health and/or personal relationships (Griffiths et al., 2012). Consequently, this has become an important issue of concern for many different stakeholder groups (e.g., parents, teachers, treatment providers, healthcare practitioners, policymakers, government bodies, and the gaming industry).

Despite the many positives of gaming, a small minority of individuals appear to engage in gaming to such an extent that it disrupts and compromises many areas of their everyday lives. Consequently, problematic gaming has become a topic of increasing research interest. However, there are multiple debates about terminology, with many terms being used interchangeably in the extant literature (e.g., 'excessive', 'problematic', 'disordered', 'dependent', 'compulsive', 'addictive', and 'pathological') (Griffiths et al., 2012). For the sake of consistency, the present review uses the term 'disorder(ed)', given that this is the term used in psychiatric diagnostic manuals.

This marked increase in research from many different perspectives (e.g., epidemiological, clinical, developmental, neurobiological, etc.), led the American Psychiatric Association (APA) to introduce 'internet gaming disorder' (IGD), as a tentative disorder in the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (APA, 2013). The APA described IGD as a behavioural addiction like gambling disorder, defining it as "persistent and recurrent use of the internet to engage in games, often with other players, leading to clinically significant impairment or distress" (p. 795) (APA, 2013). More recently, the World Health Organization (WHO) included 'gaming disorder' (GD) as a formal diagnosis in the 11th revision of the International Classification of Diseases (ICD-11) in 2019 (WHO, 2019). The criteria for both of these are shown in Table 3.

Current state of knowledge

Prevalence of gaming disorder

In the past three decades, many studies have attempted to determine the prevalence of disordered gaming. However, given the existing various definitions, screening instruments, and/or self-selected samples used, there has been a varied number of prevalence estimates across studies. To date, three meta-analyses have been published.

Fam (2018) examined the prevalence estimates of IGD among adolescents in 28 studies (N=61,737; 20 studies in Europe, four in Australia; two in Asia, and one in North America). There was wide variability is prevalence rates (0.5% to 19.9%) with a pooled prevalence rate of 4.6% of GD among adolescents (with male adolescents having higher GD prevalence rates [6.8%] than female adolescents [1.3%]). A meta-analysis by Stevens et al. (2021) comprised 53 studies (N=226,247; 17 countries). The prevalence of GD was 3.05% but lower in high quality studies (1.96%). Males had a higher GD prevalence rate (6.31%) than females (2.54%).

The most recent meta-analysis by Kim et al. (2022) comprised 61 studies (N=227,665; 29 countries). The prevalence rate of GD was 3.3% but lower when only including data from 28 representative samples (2.4%). Males had a higher GD prevalence rate (8.5%) than females (3.5%). The study also estimated prevalence rates for six different age categories. The pooled prevalence rates were 6.6% for children and adolescents (based on five studies), 6.3% for adolescents and young adults (five studies), 3.4% for young adults (nine studies), 3.3% for adolescents (38 studies), 1.9% for all adults (six studies), and 1.3% for adolescents and adults (five studies).

All three of the meta-analyses reported high heterogeneity in their reported GD prevalence rates. These were influenced by both methodological variables (e.g., screening instrument used, terminology regarding problematic gaming use, study design, type of sample surveyed, type of sampling method used) and participant variables (e.g., sample size, country/region of participants, age of participants).

Table 3. Definitions and criteria for internet gaming disorder and gaming disorder as proposed in the DSM-5 and ICD-11

	DSM-5 internet gaming disorder	**ICD-11 gaming disorder**
Definition	"Persistent and recurrent use of the internet to engage in games, often with other players, leading to clinically significant impairment or distress." (Also includes non-internet computerized games as well as internet games).	"The behavior pattern is of sufficient severity to result in significant impairment in personal, family, social, educational, occupational or other important areas of functioning."
Criteria endorsement and duration of the condition	An individual should endorse five (or more) out of nine criteria over a 12-month period.	An individual should endorse all the criteria over a 12-month period or more, although the required duration may be shortened if all diagnostic requirements are met and symptoms are severe.
Criteria	Being excessively preoccupied with gaming	Impaired control over gaming
	Having withdrawal symptoms when not gaming	Elevated priority given to gaming
	Spending more and more time gaming	Increased time spent on gaming despite problems
	Failed attempts to reduce or quit gaming	
	Losing interest in hobbies due to gaming	
	Engaging in gaming despite its adverse consequences	
	Deceiving others about gaming duration	
	Achieving a positive mood by gaming	
	Risking, jeopardizing, or losing a job or relationship because of gaming	

Note. DSM-5-TR: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision; ICD-11: International Classification of Diseases 11th Revision.

Source: Author's own elaboration

Aetiology of gaming disorder

One of the key topics in the GD field is aetiology. A recent comprehensive review on the aetiology of GD (Király et al., 2023) outlined the three over-arching interacting factors that are involved in the acquisition, development, and maintenance of GD. These are the: (i) individual factors (i.e., person-based characteristics such as genetic/biological predispositions, personality factors, motivations for playing, etc.), (ii) gaming-related factors (e.g., structural characteristics of the videogames themselves, the medium in which the videogames are played), and (iii) environmental factors (i.e., the situational characteristics such as peer, family and cultural influences in videogame playing) (Király et al., 2023).

Individual factors

Individual factors that play a contributory role in the aetiology of GD (among others) include genetic/biological predispositions, personality traits, demographic risk factors, motivations, and comorbid psychopathologies. The neural mechanisms associated with GD appear to resemble those

of other addictions (von Deneen et al., 2022). The cognitive-affective alterations found in GD include impaired executive functioning, impaired emotional regulation, impaired decision-making, and impulsivity related to different functioning in prefrontal areas and the front-limbic, temporoparietal, and subcortical regions (Schettler et al., 2022), as well structural changes in several brain regions including altered white-matter density and reduced grey matter volume (controlling emotional regulation, cognitive/motor control, decision-making, and behavioural inhibition). Studies have also indicated that compared to controls, those with GD show activation in the (i) orbitofrontal cortex (indicating lower level of punishment sensitivity), an (ii) dorsolateral prefrontal cortex (associated with higher level of craving) (Cho et al., 2022).

Many studies have explored the association between GD and the "Big Five" personality traits. Two meta-analyses have been published (Akbari et al., 2021; Chew, 2022). These have shown a very consistent positive relationship between GD and neuroticism. Given that neurotic individuals are more prone to depression, stress, and anxiety, they may use gaming as an escape because virtual worlds feel (or are perceived as) safer than their real-life personal environments. The meta-analyses also found negative associations with conscientiousness. Given that individuals with low conscientiousness are more careless, impulsive, and disorganized, the finding that they are more likely to experience GD is unsurprising. Another very consistent finding in the literature is the positive relationship between GD and impulsivity. One recent systematic literature review reported a positive relationship between impulsivity in 32 out of 33 studies (Şalvarlı & Griffiths, 2022).

As noted in the meta-analyses concerning the prevalence of GD, the literature has consistently shown that males are much more likely to experience GD than females and this also includes children and adolescents. Age also seems to be important, with adolescents and emerging adults being at higher risk of GD than other age cohorts. Various explanations have been provided in the literature from cultural perspectives (e.g., males have both a much greater affinity to, and enjoyment of, playing videogames), evolutionary perspectives (e.g., males have a greater inclination towards competition, aggression), and neurobiological perspectives (e.g., males demonstrating higher cue-elicited craving-related neural responses). Other factors have been examined but are less consistent and/or contradictory in findings related to increased risk of GD (e.g., ethnicity, relationship status, employment status, income, educational level, etc.) (Király et al., 2023).

In the case of GD, comorbidity tends to be the norm rather than the exception (Király et al., 2023). Research has consistently found a positive association between GD and (i) depression (Ostinelli et al., 2021), (ii) anxiety (Wang et al., 2017), (iii) ADHD (hyperactivity and inattention) (Király et al., 2023), (iv) comorbid polysubstance use (Burleigh et al., 2019), (v) autism (Murray et al, 2021), and (v) risk for suicidal ideation (Király et al., 2023). However, given that the majority of studies examining these associations with comorbid conditions are cross-sectional, longitudinal research is needed because the directions of the associations are uncertain. However, in many, the associations may well be reciprocal.

Gaming-related factors

To facilitate habitual and rewarding videogame playing, game design plays a role in exploiting psychological mechanisms (e.g., operant conditioning) (Király et al., 2023). For vulnerable and susceptible individuals (such as those who experience social anxiety or who have low self-esteem), such design features may facilitate excessive and (and among a minority of individuals) disordered gaming. GD (Stavropoulos et al., 2022). For instance, although GD has been reported among offline gamers, it is much more prevalent among online gamers (Király et al., 2023). Among adolescents who are socially anxious and/or who have poor social skills, online gaming environments can help meet their social needs if they find face-to-face interactions anxiety-inducing.

The genre of videogame may also contribute to GD. For instance, research has consistently found that massively multiplayer online role-playing games (MMORPGs) are most associated with GD. Other genres have been associated with GD including multiplayer online battle arena (MOBA) games, real-time strategy (RTS) games, and shooter games (both first-person and third-person (Király et al., 2023). These types of games tend to be far more immersive than other game genres and appear to be an important factor in the maintenance of GD.

The structural characteristics of the videogames themselves may also contribute to GD. Given that the virtual in-game rewards can result in the release of dopamine (Király et al., 2023), such features are critical in reinforcement and game continuance (King et al., 2010; Király et al., 2023). The unpredictability of when a reward will occur, particularly in videogames such as MMORPGs, can result in individuals playing for hours and hours in single gaming sessions. Game designers can exploit the principles of operant conditioning and players can find themselves locked into variable-ratio reinforcement schedules, which result in habitual gaming patterns.

Players designing their own in-game avatars can create extensions of themselves which may be psychologically rewarding and/or act as a compensatory mechanism for those with low body satisfaction to overcome their social anxiety, and thus boost their self-esteem (Szolin et al., 2022). Complimenting this, GD has been shown to increase when gamers experience their avatars as themselves (i.e., identification), their avatar's needs as their own (often prioritized to their offline needs [immersion]), their avatar being able to behave in ways that they cannot in their real lives (i.e., repression), and their avatar as the person/character they would like to have been (i.e., idealization) (Stavropoulos et al., 2022).

Research in media psychology-inspired concepts further reinforces the significance of structural game features for GD (Stavropoulos et al., 2022; Szolin et al., 2022). These refer to the extent gamers are absorbed by (i) the virtual world, experiencing the latter as real (i.e., as if they were there [presence/ telepresence]) (Stavropoulos et al., 2022), and (ii) their in-game activity, due the gradual increase of in-game challenges, at a rate that matches the increase of the player's in-game skills. For gamers to be challenged and completely engaged with their in-game action, these challenges need to slightly exceed their current skill level. If game-demands are significantly higher than players' skills, gamers become distressed and disengage. Similarly, if players' demands are significantly lower than their skills, they will experience boredom and disengage. As players keep engaging with the game, their skills concurrently increase, requiring the game-developer to increase the level of game challenges at a similar pace (i.e., level-up process) to maintain sustained game content consumption and process/state ("flow") (Stavropoulos et al., 2022).

Finally, the past few years has seen the introduction of arguably 'predatory' monetization techniques by the gaming industry in the form of micro-transactions (e.g., loot boxes where players spend real money to open virtual crates or boxes to win something that might help them in the progression of the games). A number of scholars have noted the similarities between loot boxes and gambling. Given that loot boxes are available to minors, it has raised concerns that loot box buying may be a 'gateway' to gambling (Király et al., 2021, 2023). Based on the empirical research to date, there appears to be a consensus that loot box buying and expenditure is indeed associated with both problematic gaming and problematic gambling among both adolescents and adults (Gibson et al., 2022; Király et al., 2023).

Environmental factors

Excluding cultural factors, research has consistently shown that early life experiences (e.g., familial relationships) can be risk factors for acquiring GD. Systematic reviews examining family factors

associated with GD among adolescents have consistently shown that specific factors in relationship quality (e.g., single parent families, family/marital conflicts, poor family functioning, poor parenting styles [neglectful, authoritarian, permissive], childhood maltreatment, violent disciplining, etc.) are positively associated with GD severity (Schneider et al., 2017; Nielsen, 2020). Other environmental factors that have been associated with adolescent GD include having difficulty in making friends, having low levels of school-related well-being (Király et al., 2023; Rehbein et al., 2013).

Future research on gaming disorder

Despite the marked increase in research examining GD and given that the majority of studies have used cross-sectional convenience sampling, further research is needed with large-scale representative samples using longitudinal designs. Also, more cross-cultural comparisons are needed - especially between Southeast Asia and Europe given the large cultural differences in these regions and variances in how parents and policymakers view gaming in the countries within them.

Further research is additionally needed from a neurobiological perspective, including whether GD may be influenced by inherited biological and/or genetic factors. There is also a dearth of data concerning clinical samples given the large reliance on community sample data. There also needs to be research into the growing area of esports (i.e., professional gaming) because playing videogames professionally can take up lots of time and resources if adolescents have aspirations to have a career in gaming (Czakó et al., 2023).

Research is also needed to help design a taxonomy relevant to current videogames and that contributes to identify which structural characteristics and game mechanics affect the behaviours of the players, especially because some of these characteristics may have age-sensitive effects. The impact of loot-box buying, for example, may be more detrimental to adolescents than adults. Finally, those in the field could also collaborate with cognate areas (such as the gambling disorder field) and try to acquire datasets from gaming operators, with the goal of identifying online gaming profiles using behavioural tracking data (e.g., using tidy classification algorithms to predict GD risk, based on engagement game mechanics [presence, flow, user-avatar bond]).

Recommendations

- One of the limitations in the field of GD field is the lack of screening instruments specifically developed for use within child and adolescent populations. Although there are a few psychometric instruments (e.g., Gaming Addiction Scale for Adolescents [Lemmens et al., 2009], Videogame Addiction Scale for Children [Yilmaz et al., 2017]) most of the screens were developed and validated with adult samples. More recent screens have relied on DSM-5 and ICD-11 criteria, which are arguably designed for adults. Therefore, bespoke age-appropriate screening instruments are vital and needed in terms of both research integrity and best clinical practices.
- As with other consumptive products that can cause problems when engaged in excessively (e.g., alcohol, gambling), there should be independent regulators in each country that oversee the videogame industry, to ensure that player protection and harm-minimization are dedicated core components of their commercial practices and goals.

- Unlike gambling and alcohol use which are adult-only activities, gaming is freely available to children and adolescents, therefore social responsibility initiatives for players need to be introduced in the same way that has happened in the gambling industry (e.g., limit setting, mandatory breaks, real-time personalized feedback, pop-up messaging on-screen, etc.) (Griffiths & Pontes, 2020).
- Research, educational awareness (for schools, parents, teachers), prevention programs, and treatment interventions should be funded by the gaming industry.
- Governments could also oblige the gaming industry to share behavioural data for research purposes.
- Countries could introduce a levy where (say) 1% of all profits are donated to an independent body for closely monitored, legitimately, and inclusively distributed funding towards these aforementioned areas and initiatives.

References

Akbari, M., Seydavi, M., Spada, M. M., Mohammadkhani, S., Jamshidi, S., Jamaloo, A., and Ayatmehr, F., 'The Big Five personality traits and online gaming: A systematic review and meta-analysis', *Journal of Behavioral Addictions,* Vol. 10, Issue 3, 2021, pp. 611-625

American Psychiatric Association, *Diagnostic and statistical manual of mental disorders (fifth edition)*, American Psychiatric Publishing, 2013

Burleigh, T. L., Griffiths, M. D., Sumich, A., Stavropoulos, V., and Kuss, D. J., 'A systematic review of the co-occurrence of gaming disorder and other potentially addictive behaviors', *Current Addiction Reports*, Vol. 6, 2019, pp. 383-401

Chew, P. K., 'A meta-analytic review of Internet gaming disorder and the Big Five personality factors', *Addictive Behaviors*, Vol. 126, 107193, 2022

Cho, T. H., Nah, Y., Park, S. H., and Han, S., 'Prefrontal cortical activation in Internet Gaming Disorder Scale high scorers during actual real-time internet gaming: A preliminary study using fNIRS', *Journal of Behavioral Addictions,* Vol. 11, 2022, pp. 492-505

Czakó, A., Király, O., Koncz, P., Yu, S. M., Mangat, H. S., Glynn, J. A., ... and Demetrovics, Z., 'Safer esports for players, spectators, and bettors: Issues, challenges, and policy recommendations', *Journal of Behavioral Addictions*, Vol. 12, Issue 1, 2023, pp. 1-8

Fam, J. Y., 'Prevalence of internet gaming disorder in adolescents: A meta-analysis across three decades', *Scandinavian Journal of Psychology*, Vol. 59, Issue 5, 2018, pp 524-531

Gibson, E., Griffiths, M. D., Calado, F., and Harris, A., 'The relationship between videogame micro-transactions and problem gaming and gambling: A systematic review', *Computers in Human Behavior*, Vol. 131, 107219, 2022.

Griffiths, M., J Kuss, D., and L King, D. 'Video game addiction: Past, present and future', *Current Psychiatry Reviews*, Vol. 8, Issue 4, 2012, pp. 308-318

Griffiths, M. D., and Pontes, H. M., 'The future of gaming disorder research and player protection: What role should the video gaming industry and researchers play?', *International Journal of Mental Health and Addiction*, Vol. 18, 2020, pp. 784-790

Kim, H. S., Son, G., Roh, E. B., Ahn, W. Y., Kim, J., Shin, S. H., ... and Choi, K. H., 'Prevalence of gaming disorder: A meta-analysis', *Addictive Behaviors*, Vol. 126, 107183, 2022.

King, D., Delfabbro, P., and Griffiths, M., 'Video game structural characteristics: A new psychological taxonomy', *International Journal of Mental Health and Addiction*, Vol. 8, 2010, pp. 90-106

Király, O., Koncz, P., Griffiths, M. D., and Demetrovics, Z., 'Gaming disorder: A summary of its characteristics and aetiology', *Comprehensive Psychiatry*, Vol. 122, 152376, 2023.

Király, O., Zhang, J., Demetrovics, Z., and Browne, D. T., 'Gambling features and monetization in video games create challenges for young people, families, and clinicians', *Journal of the American Academy of Child and Adolescent Psychiatry,* Vol. 61, 2022, pp. 854-856

Lemmens, J. S., Valkenburg, P. M., and Peter, J., 'Development and validation of a game addiction scale for adolescents', *Media Psychology*, Vol. 12, Issue 1, 2009, pp. 77-95

Lopez-Fernandez, O., Männikkö, N., Kääriäinen, M., Griffiths, M. D., and Kuss, D. J., 'Mobile gaming and problematic smartphone use: A comparative study between Belgium and Finland', *Journal of Behavioral Addictions*, Vol. 7, Issue 1, 2018, pp. 88-99

Murray, A., Koronczai, B., Király, O., Griffiths, M. D., Mannion, A., Leader, G., and Demetrovics, Z., 'Autism, problematic internet use and gaming disorder: A systematic review', *Review Journal of Autism and Developmental Disorders*, Vol. 9, 2022, pp. 120-140

Nielsen, P., Favez, N., and Rigter, H., 'Parental and family factors associated with problematic gaming and problematic internet use in adolescents: A systematic literature review', *Current Addiction Reports*, Vol. 7, 2020, pp. 365-386

Ostinelli, E. G., Zangani, C., Giordano, B., Maestri, D., Gambini, O., D'Agostino, A., ... and Purgato, M., 'Depressive symptoms and depression in individuals with internet gaming disorder: A systematic review and meta-analysis', *Journal of Affective Disorders*, Vol. 284, 2021, 136-142

Rehbein, F., and Baier, D., 'Family-, media-, and school-related risk factors of video game addiction', *Journal of Media Psychology,* Vol. 25, 2013, pp. 118-128

Şalvarlı, Ş. İ., & Griffiths, M. D., 'The association between internet gaming disorder and impulsivity: A systematic review of literature', *International Journal of Mental Health and Addiction*, Vol. 20, 2022, pp. 92-118

Schettler, L., Thomasius, R., and Paschke, K., 'Neural correlates of problematic gaming in adolescents: A systematic review of structural and functional magnetic resonance imaging studies', *Addiction Biology*, Vol., 27, Issue 1, e13093, 2022.

Schneider, L. A., King, D. L., and Delfabbro, P. H., 'Family factors in adolescent problematic Internet gaming: A systematic review', *Journal of Behavioral Addictions*, Vol. 6, Issue 3, 2017, pp. 321-333

Stavropoulos, V., Motti-Stefanidi, F., and Griffiths, M. D., 'Being young in the digital era: mental health risks and opportunities', *European Psychologist*, Vol. 27, Issue 2, 2022, pp. 86-101

Stevens, M. W., Dorstyn, D., Delfabbro, P. H., & King, D. L., 'Global prevalence of gaming disorder: A systematic review and meta-analysis', *Australian & New Zealand Journal of Psychiatry*, Vol. 55, Issue 6, 2021, pp. 553-568

Szolin, K., Kuss, D., Nuyens, F., and Griffiths, M., 'Gaming disorder: A systematic review exploring the user-avatar relationship in videogames', *Computers in Human Behavior*, Vol. 128, 107124, 2022.

von Deneen, K. M., Hussain, H., Waheed, J., Xinwen, W., Yu, D., and Yuan, K.. 'Comparison of frontostriatal circuits in adolescent nicotine addiction and internet gaming disorder', *Journal of Behavioral Addictions*, Vol. 11, Issue 1, 2022, pp. 26-39

Wang, C. Y., Wu, Y. C., Su, C. H., Lin, P. C., Ko, C. H., and Yen, J. Y., 'Association between Internet gaming disorder and generalized anxiety disorder', *Journal of Behavioral Addictions*, Vol. 6, Issue 4, 2017, pp. 564-571

World Health Organization (2019). *Gaming disorder*. https://www.who.int/standards/classifications/frequently-asked-questions/gaming-disorder

Yılmaz, E., Griffiths, M. D., and Kan, A., 'Development and validation of videogame addiction scale for children (VASC)', *International Journal of Mental Health and Addiction*, Vol., 15, 2017, pp. 869-882

3. Unpacking the Impact of Virtual Worlds

3.1. How Presence Shapes the Immersive XR Experience and Potential Wellbeing Effects

Tilo Hartmann

Department of Communication Science, Vrije Universiteit Amsterdam

Abstract

This short report addresses the basic psychology of users' immersive experiences that they make in virtual worlds, which they access via extended reality technology (XR). In contrast to established media such as video games or 2D-screen-based applications, XR technology such as Virtual Reality (VR) and Augmented Reality (AR) affords a unique immersive experience, marked by the sensation of presence. I discuss three forms presence: self-presence, where users experience their virtual body as their own; spatial presence, where the virtual environment feels as though it physically surrounds users or where virtual objects appear like physically existing here and now; and social presence, the sense of being co-located with another sentient entity. Presence is the reason why XR provides a compelling, "life-like," experience. Yet, I argue that despite this powerful immersion, users retain a degree of media awareness in XR, a cognitive recognition that the experience is technologically mediated. Both presence and media awareness shape the XR experience, and they need to be jointly considered to understand how users process information and respond to the virtual content emotionally and behaviorally. Based on this conceptualization of the XR experience, I discuss several implications for well-being, by highlighting basic potentially beneficial and problematic effects of XR. With the anticipated rise of virtual worlds that are enabled by XR, understanding the psychological fabric of the immersive experience is key to explaining how engagement in these worlds affects wellbeing.

Highlights

- The sensation of (self, spatial, social) presence represents the hallmark of the XR experience.
- Despite compelling sensations of presence, XR does not equate reality, if users are "media aware".
- Systematic research pending, the XR experience links to positive and negative effects on well-being.

Immersive or Extended Reality (XR)-Technology

What is so special about XR? One answer is to look at the underlying technology. XR-technology (or immersive technology, both terms are interchangeable) offers computer-generated sensory stimulation (visual, audio, haptic or tactile, and perhaps in the future also olfactory stimulation) that is tailored to a person's body movement (e.g., head movement, grasping, walking, etc.; Cummings & Bailenson, 2016; Slater & Sanchez-Vives, 2016). Central to XR, therefore, is that the technology tracks a user's body or motoric movement, and provides fitting real-time sensory output generated

by a computer. “Fitting real-time sensory output” implies that the provided sensory output matches the sensation a person predicted or expected to follow from a certain body movement, based on what the person commonly experienced in the actual world (Haans & Ijsselsteijn, 2012; Hartmann, 2025). While other interactive media technologies also link body movement to sensory output (e.g., a mouse cursor following movement of the hand that is steering a computer mouse), only in XR this coupling creates a real sense of “virtual reality”.

The reason is that XR technologies, such as headsets that are offering VR or AR experiences, provide a more immersive, i.e., sensory-rich, experience than traditional devices (M. Slater & Wilbur, 1997; Steuer, 1992). Present XR technologies more permanently and fully cover everything the user sees and hears, thus effectively shielding the user from receiving audiovisual input from the actual world, and more thoroughly substituting this omitted real-world data with computer-generated sensory data. Second, XR-technologies potentially provide a more natural and intuitively usable interface than previous and more artificial devices, like computer mouses, joysticks, or hand trackers that were linked to a 2D-screen (Lombard & Ditton, 2006). The envisioned ideal of XR-technologies is that users do not need to learn how to use XR, but can simply pick up a device and do things in XR as they always do, like using their hand to grasp an object, talking to another person, or actually walking towards a location to get there. To the extent this is fully realized (currently, it is not), XR will become, as an interface or medium, less apparent or visible to users while using XR, due to the smooth and natural coupling of their body movement and sensory experience. This “invisible interface”-effect is similar to the way humans, normally, do not recognize their body as an interface or medium between the external world and their internal simulation and perception (Lombard & Ditton, 2006; Riva & Mantovani, 2012). However, whereas XR might represent a more invisible interface than previous technologies, in contrast to how we experience our body as a medium, as I will address later, XR users arguably still know that their experience is mediated, because unlike the body, XR devices need to be switched on and off (Hartmann & Hofer, 2022). Accordingly, unless XR is permanently activated and built into the human body as it is done with cyborgs, I do not think that the technology will become totally transparent. Therefore, while XR provides a unique and compelling immersive experience, during exposure at least, users might always contextualize their sensations, as long as they know that this is a technology-induced experience.

The hallmark of the XR experience: Presence

XR is special, because its defining technological feature, i.e., real-time tracking linked to multi-sensory stimulation, fosters an intense sensation of presence (Felton & Jackson, 2022; Hartmann, 2025; Lee, 2004; Seth, 2014; Weber et al., 2021). Simply put, presence refers to users’ perceptual sensation that something is “happening here and now” or that someone or something appears to be physically “here, right now” (Heeter, 1992; Riccardi, 2019), even if users know that this is not actually true (Hartmann & Hofer, 2022; ISPR, 2001; Slater, 2018). Psychologically, the presence sensation arises, because XR technology successfully fulfills sensory-motor predictions of the human brain (Seth, 2014; Seth et al., 2012). Presence, however, breaks down if the XR-system is wrongly calibrated or provided spatial information is off, and the brain’s predictions are thus violated (Friedman et al., 2008; Liebold et al., 2016; Wirth et al., 2007). Three types of presence sensations are distinguished: self-presence, spatial presence, and social presence (Hartmann, 2025; Lee, 2004). All three types not only define the XR experience, but also are also key to a better understanding of how XR, or virtual worlds, might affect users’ well-being.

In XR, the user needs to slip into a virtual body, which is called an “avatar” and which is usually displayed to a certain extent (e.g, one can see one’s hands, or torso, or legs, or even full body if

looking into a virtual mirror; Nowak & Fox, 2018). **Self-Presence**, or *embodiment* (Kilteni et al., 2015), refers to users' momentary feeling that the virtual self (i.e., their avatar or virtual body; Fox et al., 2015) is their actual self. Self-presence is enhanced by cognitive top-down factors such as perceived similarity (between one's own and the virtual body), and bottom-up sensory factors such as congruence between visual and haptic information (e.g., synchronous movement of one's actual and virtual hand; Madary & Metzinger, 2016). More specifically, research suggests that users' embodiment into an avatar in XR affects three layers of self-presence (Ratan, 2013): proto, core, and extended. Proto self-presence effects suggest that a users' brain develops ownership over the virtual body, and that users automatically adapt their own body schema, accordingly (e.g., feeling thinner or taller or having longer arms). Core self-presence effects suggest that users affectively respond to factors of the virtual environment almost as if their own body would be exposed to these factors (e.g., stress, fear, or pain; (Gall et al., 2021). Extended self-presence effects refer to users' adapted self-concept, akin to the "Proteus Effect" (Yee & Bailenson, 2007), for example, feeling greater self-esteem and more confidence if embodying a very tall or beautiful avatar. Extended self-presence also captures the extent the virtual self defines, or is relevant to, a user's personal identity (Reinhard et al., 2020).

The fact that XR induces relatively strong levels of self-presence has important implications for the question to what extent XR use affects well-being. For example, XR allows people to become embodied into the bodies of other people (different gender, skin color, etc.), not only allowing them to perceive reality through the eyes of other people, but also consciously or implicitly changing their attitude about these other people (Banakou et al., 2020; Slater & Banakou, 2021). In addition, the possibility to escape the self-defining boundaries set by one's actual body might have a positive effect on well-being, e.g., via enjoyment (Slater et al., 2014). Furthermore, self-presence can yield positive effects on well-being if people are uncertain about their body, feel stigmatized, or suffer from a physical condition. Adopting a virtual body can provide momentary relief. This might be positive for well-being, for example, if people suffer from acute or chronic pain, while self-presence might provide pain relief (Matamala-Gomez et al., 2019). Perhaps more importantly, embodiment into different avatars (which can be tailored by a user) allows to experiment with different body shapes, outer appearance, and identities, which might serve self-expression and reduce uncertainties, e.g., among users from stigmatized groups (Freeman & Acena, 2022).

However, if such escapes into a more satisfying virtual body become repeatedly utilized to provide momentary relief (negative reinforcement), the risk increases that XR is used more compulsively, and that problem-focused coping in the real world is hindered. Accordingly, in this case, self-presence might spur addiction to XR (Barreda-Ángeles & Hartmann, 2022a; Brand, 2005). Relatedly, upon "re-entry" into their actual body, people might compare their own body to the virtual body and start feeling dissatisfied with their actual body shape and outer appearance (van der Waal et al., under review, chapter 7). Body dissatisfaction is considered a major problem, e.g., among social media users (Thai et al., 2024), and potentially, because one's own and other people's body shapes seem to physically exist "here and now," XR might enhance related effects (Behrens et al., 2023; Hartmann, 2025). The opposite effect is also possible, however, i.e., occupying a more desirable virtual body might transfer to a more positive perception of oneself in real-life (Yee & Bailenson, 2007). However, either way, it could be argued that a mismatch between the self suggested by a virtual vs. actual body might lead to adaptation problems in the real world. Although only speculative at this point, problems to well-being might arise if the actual self starts to pale and frustrate in light of the more compelling virtual self, or if the actual self-concept is updated in potentially dysfunctional (e.g., thinking of oneself as an aggressive male warrior) or erroneous (e.g., thinking to be more eloquent than one actually is) ways that conflict with the real-life environment.

Another potential threat to well-being that might arise from prolonged embodiment into a virtual avatar are enhanced sensations of depersonalisation (i.e., feeling alienated from one's own body) and derealisation (i.e., a feeling of unrealness, feeling estranged from the actual world; DPDR) after XR exposure (Madary & Metzinger, 2016; Peckmann et al., 2022). However, a recent large survey among 754 VR users by Barreda-Ángeles and Hartmann (2023) found no strong impact of contemporary VR use on self-reported DPDR symptoms. If anything, while positively associated with self-presence, DPDR symptoms occurred only sporadically and fleetingly after VR exposure, and more among younger users and among those who engaged in longer sessions. However, with longer exposures to more powerful future XR-systems, these effects might increase.

Spatial Presence, or *telepresence*, generally refers to users' sense of physical co-presence of the environment. In VR, spatial presence refers to feeling physically located in the virtual environment (Hartmann et al., 2015; Slater, 2018). In AR, spatial presence refers to virtual things feeling physically co-present in one's actual environment (Hadi & Park, 2024). Next to above-mentioned immersive XR features (Cummings & Bailenson, 2016), arousal, too, seems to heighten the sensation of spatial presence (Diemer et al., 2015; Peperkorn et al., 2015). Spatial presence and self-presence are closely related (Forster et al., 2022), and together they plausibly increase the immediate self-relevance of the (virtual) environment, because threats and opportunities start to seem "tangible" and directly relevant to well-being and physical health (Abraham & von Cramon, 2009; Mutlu, 2021). For example, in XR, proximity violations of one's close and intimate space causes more distress than in 2D-environments, because seemingly physical co-present things appear to threaten one's own body (Kim & Sung, 2024).

Social Presence refers to the sensation of being physically co-located with another living or sentient entity or mind (Cummings & Wertz, 2022; Oh et al., 2018). Social presence ranges from simple forms, like basic animacy (i.e., perceiving "the other" to be alive; Gray et al., 2007) to extended forms, like a sense of mutual awareness, shared cognitive focus, or even intimate mutual engagement (Biocca et al., 2003; Nowak, 2001). As social psychology has well documented, the co-presence of others turns a private into a social situation, and peoples' experience and behavior adapt accordingly. The same effect applies to mediated encounters. Research suggests that, in any mediated environment, users' belief about whether the other entity is human-controlled (avatar) or AI-driven (agent) impacts the depth of social influence (Blascovich, 2002; Fox et al., 2015; Nowak & Fox, 2018) - encountering other people's avatars is considered to be psychologically more compelling and influential than encountering a computer-controlled agent. Yet, due to an ever more sophisticated blend of generative AI (that allows producing "virtual humans") and XR (that allows displaying these virtual humans as seemingly physically existing beings in lifelike spaces), we can expect that in XR, both avatars and agents exert a relatively strong social influence on users, even if virtual humans or other agents might be still discounted (Lee, 2024). The social presence of both avatars and agents is plausibly a central factor affecting the well-being of XR users. Social presence can affect well-being in positive ways, e.g., if accompanied by supportive behavior, thus increasing the feeling of social companionship, togetherness, inclusion and support among users (e.g., Barreda-Ángeles & Hartmann, 2022). But social presence can also have a negative impact, for example, when the sensation -together with self- and spatial presence- enhances the intensity of harassment in XR (Schulenberg et al., 2023; Wiederhold, 2022). Both positive (Barreda-Ángeles & Hartmann, 2022b) and negative social presence experiences (Hinduja & Patchin, 2024; Massari et al., 2024) seem quite common among XR users, and of course particularly among users of social VR applications.

Understanding the full XR experience: Presence vs. Media Awareness

While sensations of presence represent the hallmark of the XR experience, during exposure, users might still be aware that they are not immersed in actual reality but in a simulation generated by technology. Users' belief or knowledge that "this is not truly happening" has been addressed as media awareness (Hartmann & Hofer, 2022), and it might run parallel to and thus contextualize, their sensations of presence. Media awareness might also qualify down-stream effects of presence, e.g., on the way users' feel, think, and act in XR. While media awareness seems particularly relevant to understand the XR experience, it arguably represents an essential part of any kind of media experience. For example, if exposed to Van Gogh's sunflower painting, observers might perceive the depicted sunflowers while they are, at the same time and due to the frame, oil, and canvas, aware of looking at a painting (Koblížek, 2017). Mediated content, or representations always offer two perceptual interpretations, the interface ("this is a painting") and the referent ("this is a vase with sunflowers"), and the overall media experience is only accurately described if both aspects are taken into account. This should also apply to the XR experience, where users feel that things are happening here and now (presence), but still know that these sensations are induced by XR-technology. Unlike in reality, and perhaps quite similar to lucid dreaming (Quaglia & Holecek, 2018), in XR, the presence experience is probably embedded into a knowing state that "this is not truly happening".

Building on this idea, the full XR experience can be deconstructed to better explain what it means to say that XR is very realistic and provides a virtual reality. Key to the immersive experience in XR is sensory information, and the way sensory information automatically triggers perceptual sensations, which are embedded into a set of beliefs. Only in actual reality, people believe that their perceptual sensations originate from the sensory information provided by the authentic environment. Reality is the environment that provides accurate (i.e., expected) and seemingly unmediated (i.e., no sign of an interface) sensory information. Only if things convey the expected sensory information, seem plausible and expectable, and if they seem unmediated due to the lack of any interface they seem actual and real. XR does not yet meet these criteria, and contemporary media only meet these criteria in very rare circumstances, e.g., if perceiving a Trompe-l'oeil from the right vintage point and when being ignorant of the fact that this is a painting. In contrast, if sensory information is conveyed via a detectable interface (canvas, screen, heavy HMD, or some tool that needs to be "activated"), media awareness is triggered and starts contextualizing the experience. Accordingly, even if XR induces a compelling sensation of presence, sensory inconsistencies (e.g., pixelated appearances, one cannot walk through an object, etc.), semantic inconsistencies (it is very unlikely or implausible that there's actually an alien in the living room), and voluntary reminders, reinforce media users' awareness (Hartmann & Hofer, 2022; Weber et al., 2021).

As a consequence, people might not experience XR as they experience actual reality. Also, people might behave in XR differently than they do in equivalent real-life situations. This has several implications for well-being. For example, XR, despite feeling very real, might represent a safer and more playful space than real-life (Mutlu, 2021), thus inviting users to engage in more exploratory or daring (yet, due to the high realism, also informative) behavior. For example, XR might foster greater self-disclosure in social encounters, because users feel anonymous and protected, while still offering a richer social environment than for example anonymous text-based communication (Baccon et al., 2019). However, precisely the mix of realistic presence-sensations and media awareness might also trigger dysfunctional exploratory or risky and harmful behavior, e.g., if users become curious to try out, or enjoy the experience of, harassing others.

Conclusion

In summary, XR use might affect people's mental health and well-being in many ways, positively and negatively. In general, understanding the immersive experience generated by XR allows us to better define potential effects of XR on well-being, and also the present explanation highlights several relevant issues that deserve further scrutiny (e.g., self-presence or embodiment effects on wellbeing). Yet, XR entails a broad spectrum of applications. Therefore, perhaps it is unreasonable to imagine a general effect of XR use on well-being, maybe with the exception of very generic effects such as on depersonalisation and derealisation, which might follow from any XR use (Barreda-Ángeles & Hartmann, 2023; Madary & Metzinger, 2016). In this regard, examinations of how XR affects well-being might run into the same problems as the search for a general effect of Internet use on well-being (Valkenburg et al., 2022). Rather than deriving general effects from the immersive experience, well-being effects of XR use will depend both on the user (social context, personality) and usage patterns (usage history; amount of devoted time, used content; Valkenburg & Peter, 2013). Accordingly, rather than applying the present elaboration of the immersive experience to derive general effects of XR use on well-being, a more fruitful avenue is to utilize the present elaboration of the immersive XR experience to better understand how XR affects well-being in more specific constellations (e.g., among regular younger female users of Social VR applications). Related analyses might also illuminate how XR use, under these circumstances, and due to its immersive experience, affects well-being differently than alternative mediated (e.g., using traditional online social media) or non-mediated (e.g., meeting friends offline) activities that people currently pursue.

References

Abraham, A., and von Cramon, Y.D., "Reality = Relevance? Insights from Spontaneous Modulations of the Brain's Default Network When Telling Apart Reality from Fiction", *PLoS One*, Vol. 4, No. 3, 2009, pp. 1–9.

Baccon, L. A., Chiarovano, E., and MacDougall, H.G. "Virtual Reality for Teletherapy: Avatars May Combine the Benefits of Face-to-Face Communication with the Anonymity of Online Text-Based Communication", *Cyberpsychology, Behavior, and Social Networking*, Vol. 22, No. 2, 2019, pp. 158–165.

Banakou, D., Beacco, A., Neyret, S., Blasco-Oliver, M., Seinfeld, S., and Slater, M., "Virtual Body Ownership and Its Consequences for Implicit Racial Bias Are Dependent on Social Context", *Royal Society Open Science*, Vol. 7, No. 12, 2020, p. 201848.

Barreda-Ángeles, M., and Hartmann, T., "Experiences of Depersonalization/Derealization Among Users of Virtual Reality Applications: A Cross-Sectional Survey", *Cyberpsychology, Behavior, and Social Networking*, Vol. 26, No. 1, 2023, pp. 22–27.

———, "Hooked on the Metaverse? Exploring the Prevalence of Addiction to Virtual Reality Applications", *Frontiers in Virtual Reality*, Vol. 3, 2022, p. 1031697.

———, "Psychological Benefits of Using Social Virtual Reality Platforms during the Covid-19 Pandemic: The Role of Social and Spatial Presence", *Computers in Human Behavior*, Vol. 127, 2022, p. 107047, https://doi.org/10.1016/j.chb.2021.107047.

Behrens, S. C., Tesch, J., Sun, P.J.B, Starke, S., Black, M.J., Schneider, H., Pruccoli, J., Zipfel, S., and Giel, K.E. "Virtual Reality Exposure to a Healthy Weight Body Is a Promising Adjunct Treatment for Anorexia Nervosa", *Psychotherapy and Psychosomatics*, Vol. 92, No. 3, 2023, pp. 170–179.

Biocca, F., Harms, C., and Burgoon, J.K., "Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria", *Presence: Teleoperators and Virtual Environments*, Vol. 12, No. 5, October 2003, pp. 456–480.

Blascovich, J., "A Theoretical Model of Social Influence for Increasing the Utility of Collaborative Virtual Environments", *Proceedings of the 4th International Conference on Collaborative Virtual Environments*, 2002, pp. 25–30.

Brand, J. L., "'Integration of the Cognitive and the Psychodynamic Unconscious': Comment.", *American Psychologist*, Vol. 50, No. 9, 2005, pp. 799–800.

Cummings, J.J., and Bailenson, J. N. "How Immersive Is Enough? A Meta-Analysis of the Effect of Immersive Technology on User Presence", *Media Psychology*, Vol. 19, No. 2, April 2, 2016, pp. 272–309.

Cummings, J.J., and Wertz, E.E., "Capturing Social Presence: Concept Explication through an Empirical Analysis of Social Presence Measures", Edited by Adam Joinson, *Journal of Computer-Mediated Communication*, Vol. 28, No. 1, November 4, 2022, p. zmac027.

Diemer, J., Alpers, G.W., Shiban, Y., Peperkorn, H. M., and Mühlberger, A., "The Impact of Perception and Presence on Emotional Reactions: A Review of Research in Virtual Reality", *Frontiers in Psychology*, Vol. 6, No. January, 2015, pp. 1–9.

Felton, W. M., and Jackson, R.E., "Presence: A Review", *International Journal of Human–Computer Interaction,* Vol. 38, No. 1, January 2, 2022, pp. 1–18.

Forster, P. P., Karimpur, H., and Fiehler, K., "Why We Should Rethink Our Approach to Embodiment and Presence", *Frontiers in Virtual Reality*, Vol. 3, July 4, 2022, p. 838369.

Fox, J., Ahn, S. J. G., Janssen, J. H., Yeykelis, L., Segovia, K. Y., and Bailenson, J. N., "Avatars versus Agents: A Meta-Analysis Quantifying the Effect of Agency on Social Influence", *Human-Computer Interaction*, Vol. 30, No. 5, 2015, pp. 401–432.

Freeman, G., and Acena, D., "'Acting Out' Queer Identity: The Embodied Visibility in Social Virtual Reality", *Proceedings of the ACM on Human-Computer Interaction*, Vol. 6, No. CSCW2, November 7, 2022, pp. 1–32.

Friedman, D., Brogni, A., Slater, M., Widenfeld, H. R., Antley, A., and Garau, M., "Temporal and Spatial Variations in Presence: Qualitative Analysis of Interviews from an Experiment on Breaks in Presence", *Presence: Teleoperators and Virtual Environments*, Vol. 17, No. 3, 2008, pp. 293–309.

Gall, D., Roth, D., Stauffert, J. P., Zarges, J., and Latoschik, M. E., "Embodiment in Virtual Reality Intensifies Emotional Responses to Virtual Stimuli", *Frontiers in Psychology*, Vol. 12, September 6, 2021. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.674179/full .

Gray, H. M., Gray, K., and Wegner, D. M., "Dimensions of Mind Perception", *Science*, Vol. 315, No. 5812, 2007, p. 619.

Haans, A., and Ijsselsteijn, W. A., "Embodiment and Telepresence: Toward a Comprehensive Theoretical Framework", *Interacting with Computers*, Vol. 24, No. 4, 2012, pp. 211–218.

Hadi, R., and Park, E. S., "Bridging the Digital and Physical: The Psychology of Augmented Reality", *Current Opinion in Psychology*, Vol. 58, August 2024, p. 101842.

Hartmann, T., "Being Present in Virtual Reality and Augmented Reality: Explicating the Psychology of Immersive Technology (XR)", in R Bailey and Glenna Read (eds.), *De Gruyter Handbook of Media Psychology*, de Grutyer, Berlin, 2025.

Hartmann, T., and Hofer, M., "I Know It Is Not Real (and That Matters): Media Awareness vs. Presence in a Parallel Processing Account of the VR Experience", *Frontiers in Virtual Reality*, Vol. 3, 2022, p. 694048.

Hartmann, T., Wirth, W., Vorderer, P., Klimmt, C., Schramm, H., and Böcking, S., "Spatial Presence Theory: State of the Art and Challenges Ahead", in Lombard, M., Biocca, F., Freeman, J., IJsselsteijn, W., and Schaevitz, R. J. (eds.), *Immersed in Media*, Springer International Publishing, Cham, 2015, pp. 115–135. https://link.springer.com/10.1007/978-3-319-10190-3_7

Heeter, C., "Being There: The Subjective Experinece of Presence", *Presence: Teleoperators and Virtual Environments*, Vol. 1, Issue 2, 1992, pp. 262–271.

Hinduja, S., and Patchin, J. W., "Metaverse Risks and Harms among US Youth: Experiences, Gender Differences, and Prevention and Response Measures", *New Media & Society*, 2024, p. 14614448241284413.

ISPR, *"Presence Defined",* 2001. https://ispr.info/about-presence-2/about-presence/

Kilteni, K., Maselli, A., Kording, K. P. and Slater, M., "Over My Fake Body: Body Ownership Illusions for Studying the Multisensory Basis of Own-Body Perception", *Frontiers in Human Neuroscience*, Vol. 9, 2015. https://www.frontiersin.org/articles/10.3389/fnhum.2015.00141

Kim, I., and Sung, J., "New Proxemics in New Space: Proxemics in VR", *Virtual Reality*, Vol. 28, Issue 2, March 27, 2024, p. 85.

Koblížek, T., ed., *The Aesthetic Illusion in Literature and the Arts*, Bloomsbury Academic, 2017.

Lee, E.-J., "Minding the Source: Toward an Integrative Theory of Human–Machine Communication", *Human Communication Research*, Vol. 50, Issue 2, 2024, pp. 184–193.

Lee, K. M., "Presence, Explicated", *Communication Theory*, Vol. 14, Issue. 1, 2004, pp. 27–50.

Liebold, B., Brill, M., Pietschmann, D., Schwab, F., and Ohler, P. "Continuous Measurement of Breaks in Presence", *Media Psychology*, Vol. 20, Issue 3, 2016, pp. 477–501.

Lombard, M., and Ditton, T., "At the Heart of It All: The Concept of Presence", *Journal of Computer-Mediated Communication*, Vol. 3, Issue 2, 2006, https://academic.oup.com/jcmc/article/4080403

Madary, M., and Metzinger, T. K. "Recommendations for Good Scientific Practice and the Consumers of VR-Technology", *Frontiers in Robotics and AI*, Vol. 3, Issue February, 2016, pp. 1–23.

Massari, F., Van Belle, J.-P., and Turpin, M., "Navigating New Realities: Experiences of Early Adopters in the Metaverse", *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*, ACM, Arenzano, Genoa Italy, 2024, pp. 1–3. https://dl.acm.org/doi/10.1145/3656650.3656702

Matamala-Gomez, M., Donegan, T., Bottiroli, S., Sandrini, G., Sanchez-Vives, M. V. and Tassorelli, C., "Immersive Virtual Reality and Virtual Embodiment for Pain Relief", *Frontiers in Human Neuroscience*, Vol. 13, 2019, p. 279.

Mutlu, B., "The Virtual and the Physical: Two Frames of Mind", iScience, Vol. 24, Issue 2, 2021, p. 101965.

Nowak, K., "Defining and Differentiating Copresence, Social Presence and Presence as Transportation", *Presence 2001 Conference*, Philadelphia, PA, 2001.

Nowak, K. L., and Fox, J. "Avatars and Computer-Mediated Communication: A Review of the Definitions, Uses, and Effects of Digital Representations", *Review of Communication Research*, Vol. 6, 2018, pp. 30–53.

Oh, C. S., Bailenson, J. N. and Welch, G. F., "A Systematic Review of Social Presence: Definition, Antecedents, and Implications", *Frontiers in Robotics and AI*, Vol. 5, 2018, p. 114.

Peckmann, C., Kannen, K., Pensel, M. C., Lux, S., Philipsen, A., and Braun, N. "Virtual Reality Induces Symptoms of Depersonalization and Derealization: A Longitudinal Randomised Control Trial", *Computers in Human Behavior*, 2022, p. 107233.

Peperkorn, H. M., Diemer, J., and Mühlberger, A., "Temporal Dynamics in the Relation between Presence and Fear in Virtual Reality", *Computers in Human Behavior*, Vol. 48, 2015, pp. 542–547.

Quaglia, J. T., and Holecek, A., "Lucid Virtual Dreaming: Antecedents and Consequents of Virtual Lucidity during Virtual Threat", *25th IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2018 - Proceedings*, No. March, 2018, pp. 65–72.

Ratan, R., "Self-Presence, Explicated: Body, Emotion, and Identity Extension into the Virtual Self", in R. Luppicini (ed.), *Handbook of Research on Technoself: Identity in a Technological Society*, IGI Global, 2013, pp. 322–336.

Reinhard, R., Shah, K. G., Faust-Christmann, C. A. and Lachmann, T., "Acting Your Avatar's Age: Effects of Virtual Reality Avatar Embodiment on Real Life Walking Speed", *Media Psychology*, Vol. 23, Issue 2, , 2020, pp. 293–315

Riccardi, M., "Perceptual Presence: An Attentional Account", *Synthese*, Vol. 196, Issue 7, pp. 2907–2926.

Riva, G., and Mantovani, F., "From the Body to the Tools and Back: A General Framework for Presence in Mediated Interactions", *Interacting with Computers*, Vol. 24, Issue 4, 2012, pp. 203–210.

Schulenberg, K., Freeman, G., Li, L., and Barwulor, C., "'Creepy Towards My Avatar Body, Creepy Towards My Body': How Women Experience and Manage Harassment Risks in Social Virtual Reality", *Proceedings of the ACM on Human-Computer Interaction*, Vol. 7, No. CSCW2, September 28, 2023, pp. 1–29.

Seth, A. K., "A Predictive Processing Theory of Sensorimotor Contingencies: Explaining the Puzzle of Perceptual Presence and Its Absence in Synesthesia", *Cognitive Neuroscience*, Vol. 5, Issue 2, 2014, pp. 97–118.

Seth, A. K., Suzuki, K., and Critchley, H. D., "An Interoceptive Predictive Coding Model of Conscious Presence", *Frontiers in Psychology*, Vol. 2, 2012. http://journal.frontiersin.org/article/10.3389/fpsyg.2011.00395/abstract .

Slater, M., "Immersion and the Illusion of Presence in Virtual Reality", *British Journal of Psychology*, Vol. 109, Issue 3, 2018, pp. 431–433.

Slater, M., and Banakou, D., "The Golden Rule as a Paradigm for Fostering Prosocial Behavior With Virtual Reality", *Current Directions in Psychological Science*, Vol. 30, Issue 6, pp. 503–509.

Slater, Mel, and Maria Sanchez-Vives, "Enhancing Our Lives with Immersive Virtual Reality", Frontiers in Robotics and AI, Vol. 3, No. December, 2016. http://journal.frontiersin.org/article/10.3389/frobt.2016.00074

Slater, M., and Wilbur, S., "A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments", *Presence: Teleoperators and Virtual Environments*, Vol. 6, Issue 6, 1997, pp. 603–616.

Slater, M. D., Johnson, B. K., Cohen, J., Comello, M. L. G., and Ewoldsen, D. R., "Temporarily Expanding the Boundaries of the Self: Motivations for Entering the Story World and Implications for Narrative Effects: Temporarily Expanded Boundaries of the Self", *Journal of Communication*, Vol. 64, Issue 3, 2014, pp. 439–455.

Steuer, J., “Defining Virtual Reality: Dimensions Determining Telepresence”, *Journal of Communication*, Vol. 42, Issue 4, 1992, pp. 73–93.

Thai, H., Davis, C. G., Mahboob, W., Perry, S., Adams, A., and Goldfield, G. S., “Reducing Social Media Use Improves Appearance and Weight Esteem in Youth with Emotional Distress.”, *Psychology of Popular Media*, Vol. 13, Issue 1, 2024, pp. 162–169.

Valkenburg, P. M., and Peter, J. “The Differential Susceptibility to Media Effects Model”, *Journal of Communication*, Vol. 63, Issue 2, 2013, pp. 221–243.

Valkenburg, P. M., Van Driel, I. I., and Beyens, I. “The Associations of Active and Passive Social Media Use with Well-Being: A Critical Scoping Review”, *New Media & Society*, Vol. 24, Issue 2, 2022, pp. 530–549.

Van der Waal, N. E., Janssen, L., Spapens, I.N.A., Antheunis, M., and van der Laan, L. N. Experiencing the consequences of unhealthy snacking: using qualitative methods to understand the full-body illusion of an overweight avatar in young healthy-weight adults. Submitted for publication.

Weber, S., Weibel, D. and Mast, F. W., “How to Get There When You Are There Already? Defining Presence in Virtual Reality and the Importance of Perceived Realism”, *Frontiers in Psychology*, Vol. 12, 2021, p. 628298.

Wiederhold, B. K., “Sexual Harassment in the Metaverse”, *Cyberpsychology, Behavior, and Social Networking*, Vol. 25, Issue 8, 2022, pp. 479–480.

Wirth, W., Hartmann, T., Böcking, S., Vorderer, P., Klimmt, C., Schramm, H., Saari, T., et al., “A Process Model of the Formation of Spatial Presence Experiences”, *Media Psychology*, Vol. 9, Issue 3, 2007.

Yee, N., and Bailenson, J., “The Proteus Effect: The Effect of Transformed Self-Representation on Behavior”, *Human Communication Research*, Vol. 33, Issue 3, 2007, pp. 271–290.

3.2. Novel Opportunities and Emerging Risks of Social Virtual Reality Spaces for Online Interactions

Guo Freeman

School of Computing, Clemson University

Abstract

In recent years, the growing popularity of commercial social VR platforms such as VR Chat, RecRoom, and Meta Horizon Worlds is dramatically transforming how people meet, interact, play, and collaborate online and has led to the emerging metaverse paradigm. These platforms have drawn aspects from traditional multiplayer online games and 3D virtual worlds where users engage in various immersive experiences, interactive activities, and choices through avatar-based online representations. However, social VR also demonstrates specific nuances, including full/partial body tracked avatars, synchronous voice conversations, and simulated touching and grabbing features. These novel characteristics have led to varied issues regarding people's online safety, including greater instances of online harassment and new power dynamics compared to traditional 3D virtual worlds/online gaming or single-user VR. This short report offers a comprehensive overview of novel opportunities and emerging risks of social VR for online interactions. It also highlights potential future directions for designing safer, inclusive, and more supportive social VR systems to empower diverse communities, especially marginalized users such as women, ethnic minorities, and LGBTQ individuals.

Highlights

- Social VR has offered novel opportunities for future social interactions in individual, interpersonal, and community dimensions.
- Social VR also leads to various emerging risks as it creates new power dynamics in online spaces.
- Future work should focus on designing social VR systems as inclusive, supportive, and empowering novel social spaces for diverse communities.

Introduction

Social Virtual Reality (VR) platforms are novel and increasingly popular 3D virtual spaces where multiple users can interact with one another through VR head-mounted displays (McVeigh-Schultz et al., 2018, 2019). Rather than merely looking at avatars on a computer screen, social VR provides partially or fully body-tracked avatars (i.e., one's avatar movements correspond to one's offline body movements in real-time), synchronous voice conversations, and simulated touching and grabbing features. These unique features thus allow users to socialize in more embodied (i.e., experiencing a virtual body representation as one's own; see Slater et al., 2009) and immersive (i.e.,

being enveloped by, included in, and interacting with the virtual environment; see Witmer & Singer, 1998) ways. Existing literature has shown that these unique and novel features have attracted diverse users of different age groups, genders, races, sexual orientations, and abilities, including many marginalized or historically ignored individuals in tech spaces (e.g., women, LGBTQ individuals, ethnic minorities, the otherly-abled, and people with intersectional identities) (Feeman & Acena, 2021, 2022; Freeman & Maloney, 2021; Freeman et al., 2022b; Li et al., 2023; Maloney & Freeman, 2020; Maloney et al., 2021). As social VR is playing an essential role in the emerging metaverse paradigm and becoming an entrenched norm of our virtual society, we envision that social VR will become even more critical in diverse users' networked lives by offering them novel and immersive social interactions.

However, social VR's focus on embodied and immersive experiences has also led to varied issues regarding people's online safety, including intensified and more physicalized forms of harassment in social VR compared to other online contexts, ranging from trash-talking women, drawing penises, and virtual "groping" to the most recent "rape" in the metaverse (Blackwell et al., 2019a, 2019b; Freeman et al., 2022c; Schulenberg et al., 2023a, 2023b, 2023c; Zheng et al., 2023). Therefore, it is important to comprehensively investigate both novel opportunities and emerging risks of social VR for online interactions, which is crucial for designing safer, inclusive, and more supportive social VR systems to empower diverse communities, especially marginalized users, in the future.

Novel Opportunities of Leveraging Social VR for Future Social Interactions

Over the past five years, social VR has constituted increasingly popular digital social spaces where diverse users meet, interact, and socialize in new and more immersive ways and in various contexts (e.g., conferences, workshops, meetings, camps, public events such as concerts, and classroom teaching). In social VR, people can engage in a wide range of experiences (e.g., cooking, dancing, or falling asleep with someone else), which raises crucial questions on how social VR may significantly impact every aspect of people's daily lives in the near future. In this sense, social VR is playing a critical role in the emerging metaverse paradigm where various virtual worlds, augmented reality, and the Internet are all seamlessly intertwined. There is also a growing research agenda on design strategies for future social VR spaces, such as a design framework for shaping pro-social behavior in VR (Jonas et al., 2019; McVeigh-Schultz et al., 2019), methods to design non-verbal communication in social VR (Maloney et al., 2020b, Tanenbaum et al., 2020), strategies to design valuable social VR experiences for older adults (Baker et al., 2019), and exploring "weird" forms of sociality and embodiment in VR/XR for future everyday life (McVeigh-Schultz & Isbister, 2021).

In particular, the unique socio-technical features of social VR have attracted diverse communities, which demonstrate its potential to be widely used among marginalized and historically ignored populations in the near future (McVeigh-Schultz et al., 2018, 2019; Outlaw & Duckles, 2018; Li et al., 2023). As prior work has shown, diverse users, including marginalized and historically ignored populations, generally enjoy their social VR experiences as embodied and immersive interactivity (Freeman & Acena, 2022; Freeman & Maloney, 2021; Freeman et al., 2022b; Maloney et al., 2020a, 2021). Existing works have also highlighted five forms of social activities in social VR that diverse users find subjectively meaningful and valuable: full-body "mirroring," performing mundane and essential everyday activities in new ways, activities for social and mental self-improvement, immersive cultural appreciation and educational activities, and engaging in immersive events (Maloney & Freeman, 2020).

Taken together, social VR has shown its potential to benefit diverse communities by offering novel opportunities for future social interactions in **individual**, **interpersonal**, and **community** dimensions.

In the **individual** dimension, social VR allows diverse users, especially marginalized individuals, to engage with their identity openly and establishing unique connections between their avatar and physical body (Freeman & Acena, 2022; Freeman & Maloney, 2021; Freeman et al., 2022b). For example, non-cisgender individuals have leveraged social VR platforms to present, express, and experiment their identity in ways that traditional online social spaces cannot provide, including: experimenting embodied avatars, leveraging voice chat to train and validate a gender appropriate voice, and community engagement with other non-cisgender users and supporters in an immersive way (Freeman et al., 2022b). This especially highlights how the physical body is re-introduced and re-discovered in the social VR context, which leads to new and novel phenomena and practices of approaching diverse gender identities online.

In the **interpersonal** dimension, social VR unlocks more fulfilling technology-mediated conversations and dynamics interactions with others (Maloney & Freeman, 2020; Maloney et al., 2020b; Moustafa & Steed, 2018; Sra et al., 2018), which may help form new intimate relationships. For example, prior works have focused on what makes non-verbal communication in social VR unique and socially desirable, such as: as more immersive and embodied interactions for body language; as a similar form of communication to offline face-to-face interaction in terms of spatial behavior, hand behavior, and facial expressions; and as a natural way to initiate communication with online strangers (Maloney et al., 2020b). As a result, social VR has been used to maintain long-distance couples' relationships (Zamanifard & Freeman, 2019), build close interpersonal relationships (Freeman & Acena, 2021), and facilitate remote collaborative work (Freeman et al., 2022a) in more immersive and embodied ways. This is especially valuable for communities who often lack offline social networks for social support. For instance, research has demonstrated how social VR innovates traditional online support mechanisms to empower LGBTQ+ individuals by creating a sense of co-presence similar to face-to-face interaction despite being online; simulating physical behaviors to demonstrate embodied support for LGBTQ+ individuals; and imitating offline LGBTQ+ centered events in a natural and immersive way (Li et al., 2023).

Therefore, in the **community** dimension, social VR also provides "embodied visibility" to help people build their collective visibility and a supportive community beyond geographic limitations (Freeman & Acena, 2021). For example, queer social VR users often employ three main strategies to build and experience embodied visibility in social VR: visualizing queer identity through avatar creation and design; acting out queer identity via full-body tracking and engaging in immersive embodied events; and vocalizing queer identity through voice communication. As a result, experiencing embodied visibility in social VR may help queer users build a supportive queer community beyond geographic limitations and even transform their visibility from online to offline (Freeman & Acena, 2021).

Emerging Risks of Leveraging Social VR for Future Social Interactions

Despite the above-mentioned novel opportunities, social VR also leads to various emerging risks as it creates new power dynamics in novel online spaces. Above all, despite attracting diverse users, social VR still seems to be a privileged space as popular social VR platforms are generally considered English-speaking, White, male, and cisnormativity dominated (Blackwell et al., 2019a;

Freeman & Acena, 2021; Freeman et al., 2022c). Engaging in social VR can, unfortunately, be a double-edged sword for people's online safety. To provide a more comprehensive image of these emerging risks in social VR, Zheng et al. (2023) investigated 212 YouTube videos and their transcripts that document social VR users' immediate experiences of safety risks as victims, attackers, or bystanders. They also analyzed spectators' reactions to these risks shown in comments to the videos (Zheng et al., 2023). In particular, they use the term "safety risk" in social VR to describe various types of detrimental user behaviors involving abusive communications directed towards other users (e.g., harassment, verbal abuse) and disruptive behaviors that violate the rules and social norms of the platform (e.g., griefing, spamming, and cheating) (Zheng et al., 2023). In total, they identified 5 types of emerging virtual risks that are unique in the social VR setting, including: *role-playing, fraud-impersonation, immersive dweller, misuse safety features, misread cues,* and *minors picking on adults*. They have also identified 7 categories of safety risks in social VR that are more severe than in other online contexts such as gaming, including: *virtual violence, virtual scaring, virtual abuse, virtual sexual harassment, virtual crashing, virtual voice trolling*, and *virtual trash actions* (Zheng et al., 2023).

Likewise, a growing body of works have highlighted the severe risk of "embodied harassment" in social VR. In this context, harassing behaviors are both conducted and experienced through a sense of embodiment about one's virtual body, such as a higher awareness of body ownership and more physical and transformative/ interactive experiences (Freeman et al., 2022c). These works have collectively warned that social VR's focus on embodiment, the sense of presence, body tracking, and synchronous voice conversation may allow people to virtually "touch" (e.g., handshaking, hugging, and high-fiveing) and assault others, leading to heightened harassment risks (Blackwell et al., 2019a; Freeman & Acena, 2021; Freeman et al., 2022c). Compared to traditional VR that largely focuses on single-player games or applications, harassment that occurs in social VR may be felt even more immersive and thus destructive due to social VR's focus on supporting open virtual worlds, simulating familiar social contexts (e.g., multi-user events), and attracting a broad range of users. For example, the same unique features that support women's, LGBTQ individuals', and ethnic minority users' identity practices in social VR could also lead to more complicated and severe forms of "embodied" harassment towards them.

Indeed, some prior work on social VR has pinpointed that individuals who are considered marginalized in tech spaces (e.g., women, LGBTQ, and ethnic minorities) may face additional harassment risks in social VR (e.g., Blackwell et al., 2019a; Freeman & Acena, 2021; Freeman et al., 2022b, 2022c; Schulenberg et al., 2023). For example, a technology report demonstrates the gender disparity amongst victims of sexual harassment in social VR, and how harassing comments are often racist and homophobic (Outlaw & Duckles, 2018). A 2017 technology report with 13 social VR women users reveals several safety risks for women in social VR, such as sexual harassment and flirting (Outlaw & Duckles, 2017). Freeman et al. (2022c) highlight that harassment in social VR may be felt as more disruptive to marginalized populations because it is easier to identify and target them as potential victims due to the combination of avatar design and voice in social VR. Blackwell et al.'s works also shed light on the risks of sexual harassment and stalking these populations face in social VR (Blackwell et al., 2019a, 2019b). Schulenberg et al. especially focused on women's experiences of harassment risks in social VR as compared to harassment targeting women in pre-existing, on-screen online gaming and virtual worlds, along with strategies women employ to manage harassment in social VR with varying degrees of success (Schulenberg et al., 2023). Their findings show that women social VR users often experience (1) violations of personal

physical space and abilities beyond "viewing" a 2D screen; (2) embodied sexual harassment due to a more nuanced avatar-self relationship; (3) harassment based on the comparatively ubiquitous use of voice communication; and (4) internalized shame and fear in and out of social VR compared to pre-existing online gaming and virtual worlds. Worse still, women also feel that existing safety features and strategies in social VR to protect them are also often insufficient, either because they rely on retroactive action, prevent potential positive interactions, or create emotional and interpersonal burdens (Schulenberg et al., 2023).

In addition, while “embodied visibility” in the current social VR spaces is novel and powerful, it is also privileged – it may make certain communities and populations more "invisible" than others, which does not help achieve an inclusive and supportive digital society. This dilemma thus leads to important questions such as: how can transgender users be visible given the common cisnormative expectations for their voices? (Freeman & Acena, 2022). Similar to the offline world, in social VR, certain specific queer subcultures with intersectional identities (e.g., transgender people of color) tend to be more marginalized than others (e.g., white gay men) as they may encounter intersectional challenges (e.g., both transphobia and racism). These concerns thus limit to what degree they are willing to present and express their identity in social VR, or if they are willing to do so at all, which could reinforce homonormativity and cisnormativity. Therefore, there is also a critical need to address the underlying challenges regarding who gets to be visible and/or more visible than others in emerging tech spaces.

Conclusions and Future Directions

The HBO documentary "We Met in Virtual Reality," filmed entirely in the most widely-used social Virtual Reality (VR) platform VRChat, captures the growing power and potential for social VR to enhance people’s digital lives in unique ways (HBO, 2022). As shown in this short report, social VR’s specific nuances, including full/partial body tracked avatars, synchronous voice conversations, and simulated touching and grabbing features, have led to both novel opportunities for innovating online social interactions and varied issues regarding people’s online safety, including greater instances of online harassment and new power dynamics compared to traditional 3D virtual worlds/online gaming or single-user VR.

In this sense, how to better support diverse communities, especially marginalized populations, by helping them navigate new power dynamics in emerging social VR spaces is still an ongoing and critically needed research agenda. First, while prior works provide empirical evidence on how “embodied visibility” in current social VR spaces can be both novel and privileged, little is known regarding who can be marginalized in social VR and why and how social VR introduces new power dynamics online. Second, it has been reported that social VR can be leveraged to afford social support for diverse communities. However, there is little knowledge on the mechanisms through which it can facilitate social support for diverse communities, what exactly these new forms of VR-mediated social support are, and how these novel forms of support can help diverse communities navigate emerging marginalization and power dynamics. Lastly, little is known regarding what new theories and designs, which go beyond the current focus on improving accessibility in VR environments (Jain et al., 2021; Ji et al., 2022; Mott et al., 2020), are needed to build inclusive and equitable social VR systems to empower diverse social VR users in individual (e.g., identity formation, self-confidence, mental health, and self-care), interpersonal (e.g., friendship, mutual help,

and collaboration), and community (e.g., community support, collaborative learning, gender equality, and LGBTQ rights) dimensions.

In conclusion, it is crucial for future work to focus on designing emerging social VR systems as inclusive, supportive, and empowering novel social spaces for diverse communities, especially marginalized or historically ignored populations in tech spaces. As our modern social lives have placed more focus on understanding and re-imagining virtual experiences, such research has the potential to significantly transform and benefit the modern-day digital lives of diverse communities, especially for marginalized or historically ignored populations in tech (e.g., women, LGBTQ, ethnic minorities, and people with intersectional identities), by providing them with healthier and more supportive interaction dynamics in individual, interpersonal, and community dimensions in the emergent metaverse paradigm.

References

Baker, S., Kelly, R. M., Waycott, J., Carrasco, R., Hoang, T., Batchelor, F., Ozanne, E., Dow, B., Warburton, J., and Vetere, F., 'Interrogating social virtual reality as a communication medium for older adults', *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 2019, pp. 1–24.

Blackwell, L., Ellison, N., Elliott-Deflo, N., and Schwartz, R., 'Harassment in social virtual reality: Challenges for platform governance', *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 2019a, pp. 1–25.

Blackwell, L., Ellison, N., Elliott-Deflo, N., and Schwartz, R., 'Harassment in social VR: Implications for design', *2019 IEEE Conference on Virtual Reality and 3D User Interfaces* (VR), 2019b, pp. 854–855.

Freeman, G. and Acena, D., 'Hugging from a distance: Building interpersonal relationships in social virtual reality', *ACM International Conference on Interactive Media Experiences*, 2021, pp. 84–95.

——— 'Acting out queer identity: The embodied visibility in social virtual reality', *Proceedings of the ACM on Human- Computer Interaction*, 6(CSCW2), 2022a, pp. 1–32.

Freeman, G., Acena, D., McNeese, N. J., and Schulenberg, K., 'Working together apart through embodiment: Engaging in everyday collaborative activities in social virtual reality', *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP), 2022a, pp. 1–25.

Freeman, G. and Maloney, D., 'Body, avatar, and me: The presentation and perception of self in social virtual reality', *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3), 2021, pp. 1–27.

Freeman, G., Maloney, D., Acena, D., and Barwulor, C., '(Re) discovering the physical body online: Strategies and challenges to approach non-cisgender identity in social virtual reality', *Proceedings of the 2022 CHI conference on human factors in computing systems*, 2022b, pp. 1–15.

Freeman, G., Zamanifard, S., Maloney, D., and Acena, D., 'Disturbing the peace: Experiencing and mitigating emerging harassment in social virtual reality', *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1), 2022c, pp. 1–30.

HBO (2022). *We met in virtual reality*. https://www.hbo.com/movies/we-met-in-virtual-reality

Jain, D., Junuzovic, S., Ofek, E., Sinclair, M., R. Porter, J., Yoon, C., Machanavajhala, S., and Ringel Morris, M., 'Towards sound accessibility in virtual reality', *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 80–91

Ji, T. F., Cochran, B. R., and Zhao, Y.. 'Demonstration of VRbubble: Enhancing peripheral avatar awareness for people with visual impairments in social virtual reality', *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 2022, pp. 1–6.

Jonas, M., Said, S., Yu, D., Aiello, C., Furlo, N., and Zytko, D., 'Towards a taxonomy of social VR application design', *Extended abstracts of the annual symposium on computer-human interaction in play companion extended abstracts*, 2019, pp. 437–444.

Li, L., Freeman, G., Schulenberg, K., and Acena, D., '"We cried on each other's shoulders": How LGBTQ+ individuals experience social support in social virtual reality', *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–16.

Maloney, D. and Freeman, G., 'Falling asleep together: What makes activities in social virtual reality meaningful to users', *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, 2020, pp. 510–521.

Maloney, D., Freeman, G., and Robb, A., 'A virtual space for all: Exploring children's experience in social virtual reality', *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, 2020a, pp. 472–483.

——— 'Stay connected in an immersive world: Why teenagers engage in social virtual reality', *Interaction Design and Children*, 2021, pp. 69–79.

Maloney, D., Freeman, G., and Wohn, D. Y., 'Talking without a voice: Understanding non-verbal communication in social virtual reality', *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2), 2020b, pp. 1–25.

McVeigh-Schultz, J. and Isbister, K., 'The case for "weird social" in VR/XR: A vision of social superpowers beyond meatspace', *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–10.

McVeigh-Schultz, J., Kolesnichenko, A., and Isbister, K., 'Shaping pro-social interaction in VR: An emerging design framework', *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

McVeigh-Schultz, J., Márquez Segura, E., Merrill, N., and Isbister, K., 'What's it mean to" be social" in VR? mapping the social VR design ecology', *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems,* 2018, pp. 289–294

Mott, M., Tang, J., Kane, S., Cutrell, E., and Ringel Morris, M., '"I just went into it assuming that I wouldn't be able to have the full experience": Understanding the accessibility of virtual reality for people with limited mobility', *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 2020, pp. 1–13.

Moustafa, F. and Steed, A., 'A longitudinal study of small group interaction in social virtual reality', *Proceedings of the 24th ACM symposium on virtual reality software and technology*, 2018, pp. 1–10.

Outlaw, J. and Duckles, B., 'Why women don't like social virtual reality: A study of safety, usability, and self-expression in social vr', 2017, https://www.extendedmind.io/why-women-dont-like-social-virtual-reality.

Outlaw, J. and Duckles, B., 'Virtual harassment: The social experience of 600+ regular virtual reality (VR) users' https://virtualrealitypop.com/virtual-harassment-the-social-experience-of-600-regular-virtual-reality-vr-users-23b1b4ef884e

Schulenberg, K., Freeman, G., Li, L., and Barwulor, C., 'Creepy towards my avatar body, creepy towards my body: How women experience and manage harassment risks in social virtual reality', *Proceedings of the ACM on Human-Computer Interaction*, 2023a, 7(CSCW2), pp. 1-29.

Schulenberg, K., Li, L., Freeman, G., Zamanifard, S., and McNeese, N. J., 'Towards leveraging AI-based moderation to address emergent harassment in social virtual reality', *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023b, pp. 1–17.

Schulenberg, K., Li, L., Lancaster, C., Zytko, D., and Freeman, G., '"We don't want a bird cage, we want guardrails"': Understanding designing for preventing interpersonal harm in social VR through the lens of consent', *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2), 2023c, pp. 1-30.

Shriram, K. and Schwartz, R., 'All are welcome: Using VR ethnography to explore harassment behavior in immersive social virtual reality', *2017 IEEE Virtual Reality* (VR), 2017, pp. 225–226.

Slater, M., Pérez Marcos, D., Ehrsson, H., and Sanchez-Vives, M. V., 'Inducing illusory ownership of a virtual body', *Frontiers in Neuroscience*, 2009, p. 29.

Sra, M., Mottelson, A., and Maes, P., 'Your place and mine: Designing a shared VR experience for remotely located users', *Proceedings of the 2018 Designing Interactive Systems Conference*, 2018, pp. 85–97.

Tanenbaum, T. J., Hartoonian, N., and Bryan, J., '" How do I make this thing smile?": An inventory of expressive nonverbal communication in commercial social virtual reality platforms', *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–13.

Witmer, B. G. and Singer, M. J., 'Measuring presence in virtual environments: A presence questionnaire', *Presence*, Vol. 7, Issue 3, 1998, pp. 225–240.

Zamanifard, S. and Freeman, G., 'The togetherness that we crave experiencing social VR in long distance relationships', *Conference companion publication of the 2019 on computer supported cooperative work and social computing*, 2019, pp. 438–442

Zheng, Q., Xu, S., Wang, L., Tang, Y., Salvi, R. C., Freeman, G., and Huang, Y., 'Understanding safety risks and safety design in social VR environments', *Proceedings of the ACM on Human-Computer Interaction*, Vol. 7(CSCW1), 2023, pp. 1–37

3.3. Harmful Design Practices in Virtual Worlds

Xinning Gui

The Pennsylvania State University

Abstract

As the design of virtual worlds becomes growingly democratized, virtual worlds enjoy a diverse array of user-generated content but also suffer from harmful design practices resulting in problematic content and mechanisms that have a harmful effect on the wellbeing of virtual world users. Compared to harmful content in traditional forms such as text, image, and video, harmful designs capture a dynamic process where the orchestration of innocuous elements could lead to harmful effects through their interactions with users, rendering them exponentially difficult to define and detect. The harmful effects include but are not limited to financial loss, privacy violation, and exposure to explicit or extremist content. But harmful design practices cannot be blamed on individual designers alone. An array of individual and ecological circumstances can be linked to the rise of harmful design practices, ranging from individual aspirations to success to platform-level problematic financial incentive structures. To ameliorate harmful design practices, one important consideration is to pay attention to and enhance the ethical agency in virtual world design.

Highlights

- Harmful designs denote design outcomes that have a harmful effect emerging out of user interaction.
- Harmful design practices are conditioned in both individual and ecological circumstances.
- Designers of harmful virtual worlds are both perpetrators and victims.
- Ethical agency is distributed across multiple stakeholder groups beyond virtual world designers.

Harmful design: Concept, Characteristics, and Typology

Virtual worlds provide users with immersive experiences, affording expanded opportunities for communication (Castronova, 2007), learning (Bendis, 2007), and work (Dionisio et al., 2013). However, virtual worlds are not safe from harm. While prior literature has already covered various sorts of harms, such as hate speech (Rowland, 2011), harassment (Wiederhold, 2022), and cyberbullying (Aponte & Richards, 2013), the particular type of harm that this short report focuses on is harmful design, defined broadly as design outcomes that have a harmful effect on virtual world users' wellbeing. Virtual worlds today are criticized for hosting harmful designs. A telling example in the news is that a virtual world could use harmful designs, which include strict, simple rules that exert nearly complete control over what users could wear or do, so that unwitting child users are induced to practice slavery and Nazi role-plays, and subsequently to coordinate harassment campaigns against other users (Cecilia D'Anastasio, 2021).

Harmful design is distinct from harmful content in static forms such as text, image, or video. Harmful content in these traditional forms can be directly recognized by victims or bystanders as they draw from platform policies, or general societal values and standards to assess what they have experienced or witnessed (e.g., (Cowan et al., 2002; Leopold et al., 2019)). Platforms usually provide users with a flag function to report harmful content (Kou & Gui, 2021), but harmful design can be less perceptible to affected players. A virtual world's surface content may not violate any policy, but its mechanisms that manipulate user behavior have a harmful effect. For instance, our recent interview study with virtual world designers (Kou et al., 2025) suggested that they would intentionally design colorful and jumpy characters for child users so they are more likely to extend their engagement time and spend more.

In addition, since harmful design is exponentially more complex than static forms of harmful content, the ways harm can be embedded in virtual world design are also multiple. For instance, a virtual world could be designed to present sexual content through a combination of suggestive visual elements that could bypass traditional text-based or image-based automated detection (Jargon, 2021). A secretive social place could be designed to facilitate sex-related behaviors without attracting the attention of moderation (Kou & Gui, 2023). When harm perpetrators' capacity is amplified by virtual world design, the impact and severity of harm they committed could also be amplified.

Particularly, our prior work (Kou & Gui, 2023) has identified four primary types of harmful designs:

1) **Ubiquitous microtransaction design** refers to the prevalence of purchase mechanisms embedded in users' virtual world experiences. Imagine entering a virtual world and frequently encountering features that users must pay in order to progress or collect essential items. While the existence of microtransactions is financially important for virtual worlds, especially those that are free to play, its overuse leads to obstructed user experience, misled user behavior, and possibly financial harm to unsuspecting users. What is more concerning is gambling-like designs such as roulette and loot box, the use of which among young users is linked to problem gambling behavior (Brooks & Clark, 2023; Zendle et al., 2019).

2) **Unconstrained social design** means the design of social spaces where players can have social interactions with others. While social spaces may appear innocuous, they need an extra layer of design consideration to protect the wellbeing of child users. Social spaces that are unmoderated can easily attract explicit language and behavior, as well as sexual predators.

3) **Unmoderated expression design** denotes the design of channels that mediate users' communication within a virtual world. Communication channels within a virtual world are oftentimes poorly moderated and thus susceptible to abuse and misuse. For example, in the 'virtual plaza' type of virtual worlds on Roblox, users can hold a virtual booth and decorate their booth with any content, oftentimes unmoderated by Roblox.

4) **Problematic world design** captures how the main theme of a virtual world features problematic ideologies such as slavery, Nazism, and terrorism. Those problematic themes could often have a long-lasting impact on child users who are unsuspecting. Our data has plenty of adult players who could still remember the problematic world designs they experienced as a child many years ago.

While this represents a preliminary understanding of types of harmful designs in virtual worlds, it has a generative nature in the sense that further work can be built upon it to construct a more comprehensive taxonomy, at scale.

Harmful design and harmful design practices

Harmful design is the outcome of harmful design practice. Design practice should not be taken lightly, as it denotes the professional designer's constant negotiation with ongoing challenges (e.g., limited resources and limited time) and complex and difficult solutions (Stolterman, 2008). Virtual world designers also need to negotiate with a range of factors such as their target users, budget and resources for generating a virtual world, and potential of monetizing the virtual world even to just cover the costs of development and maintenance.

Behavioral scientists and privacy and security researchers have explored design practices such as nudging (Hanna, 2015) in interface design to warn users about privacy risks (Wang et al., 2013). However, deliberate designs could also deprive people of autonomy or choices (Cave, 2006). Researchers and practitioners (Gray et al., 2018; Mathur et al., 2019) have started to pay attention to negative interaction design patterns such as deceptive patterns, which aim to advance shareholders' interests at the expense of end users' (Gray et al., 2018), incurring various harms to individual welfare, financial state, and data privacy (Bongard-Blanchy et al., 2021). For instance, mobile app interfaces could be designed in such a way to increase the difficulty for the user to find the 'unsubscribe' button, or preselect a choice that is not in the user's best interest (Gray et al., 2018). Sometimes, users may not be fully aware of the deliberate design intent, but their decision-makings and interactions could be influenced in favor of technology owners' interests (Gray et al., 2021); other times, users could be aware of manipulation, but do not know how to oppose the influence (Bongard-Blanchy et al., 2021).

If deceptive interface design patterns such as interface layout, color design, and sequence of user actions, and incur mostly time or monetary loss (Gray et al., 2018). Harmful design in virtual worlds is distinct as it involves design beyond the interface level and comes from an orchestration of elements in virtual worlds. With a whole virtual world at their disposal, the designer can maneuver a wide array of elements, such as task design, difficulty design, reward and punishment mechanisms, and avatar design, to orchestrate a desired harmful effect. Such immersiveness could impact users' cognitive processes such as rational thinking (Al-Jarani, 2019), and work in lockstep with harmful design to bypass users' rational agency.

Harmful design practices do not exist with individual designers alone but can be amplified through online communities. In our recent investigation of an online community of virtual world designers (Zhang et al., 2024), we observed a recurring theme that virtual world designers actively engage in exploring and sharing harmful design ideas with one another, potentially further compounding the risk for virtual world users. Specifically, virtual world designers openly share harmful design ideas, such as a terrorism-themed virtual world that simulates the September 11 attacks or making a casino-based game. They also explore design strategies to bypass virtual world moderation systems so that their virtual world design will not be detected, such as revising a few pixels on an image so that it can go through the automated moderation system. As such, the ideation part of harmful design practices can be a collaborative process which reproduce and perpetuate harmful design ideas.

The ecological context of harmful design practices

To analyze harmful design practices, it is insufficient to look at the isolated design practice alone and attribute it to the virtual world designer only. As our project unfolds, we traced the emergence of harmful design practices to a few deeper ecological roots, which form a much clearer picture of what is happening in harmful design practices. Our recent study (Kou et al., 2024) undertook this

task to examine how harmful designs take place in user-generated virtual worlds. Particularly, we identified three interconnected dimensions, namely sociotechnical risks, socioeconomic precarities, and normative insensitivities:

1) First, virtual world designers, especially end users who aspire to create a virtual world, are not trained to be one. They have to learn on their own through available resources such as tutorial videos on YouTube and online designer communities. As a result, inexperienced virtual world designers can be exposed to several types of risks. They may be attracted to open-source codes for generating virtual worlds but lack sufficient technical expertise to tell whether those codes contain viruses. As a result, their virtual worlds can be sabotaged by viruses, and, in worser scenarios, be removed by platforms that host their virtual worlds. Virtual world designers are also vulnerable to scammers who claim to be wanting to collaborate on a large virtual world, but only end up stealing or destroying their source code. Thus, it is insufficient to frame the designer as the perpetrator of harmful design. We need to see the larger context, where the designer themselves also has plenty of vulnerabilities.

2) Second, virtual world designers can be in a situation of socioeconomic precarity. In the age of platform economy, they design virtual worlds not simply out of passion. They have and need financial gains to maintain their virtual worlds. Platforms like Roblox run a revenue-sharing business model to share profits with virtual world designers. However, the power relation between platforms and virtual world designers are not balanced. Virtual world designers only get a small cut, and the skills they have developed are untransferable. As a result, they are coerced to stay on the same platform, and, with the profit-driven business ethos, they are financially incentivized to produce harmful designs, so long as they are lucrative.

3) Third, virtual world designers also become insensitive to the risks associated with harmful designs that they create. As they give priorities to overcoming the sociotechnical risks and socioeconomic precarities, safety of their virtual world users is naturally deprioritized. Thus, they also normalize harmful designs so long as they are lucrative, and they openly discuss strategies that can challenge and bypass moderation systems that aim to detect their harmful designs.

Overall, it has become clear that it is no longer adequate to attribute harmful design to the design practice, but instead we need to look beyond and ask what is the context that enables and legitimates such design practice. The ecological context of harmful design practices suggests that it is important to interrogate the role platforms play in the ecology of harmful design, and it is critical to seek solutions both within and beyond a single platform.

Designing for user wellbeing in virtual worlds

The solution to harmful design should involve multiple layers of attention as well as diverse stakeholder groups. However, design ethics is still critical in reflecting on harmful design practice, as well as how to design for user wellbeing. There are several helpful perspectives on this. First, it is important to acknowledge that ethical agency does not solely lie with the virtual world designer. If users are able to recognize ethical violations (Kou & Gui, 2023), then it is natural to utilize a multi-stakeholder framework to identify stakeholder groups that are impacted by unethical design decisions and that can make ethical assessments of existing harmful designs to work together to devise an ethical design framework for virtual world design. Second, value-based design frameworks such as value sensitive design (Friedman, 1996) and values at play (Flanagan et al., 2005) can be used to evaluate existing harmful designs and inform key values for user wellbeing-oriented virtual world design. Lastly, rights-based approaches (Pothong et al., 2024) can be used to

identify high-level, universal values that should hold across various virtual worlds in the world. These are far from an exhaustive list of possible design approaches but provide suitable starting points for us to consider how future virtual world design practices can reduce harm and enhance user wellbeing. We should remain hopeful as more scholarly and regulatory attention is accumulating for meaningful changes in how we design virtual worlds.

References

Al-Jarani, Y., 'All fun and (mind) games? Protecting consumers from the manipulative harms of interactive virtual reality', *University of Illinois Journal of Law, Technology & Policy*, Vol. 2, pp. 299–353

Aponte, D. F. G., and Richards, D., 'Managing cyber-bullying in online educational virtual worlds', *Proceedings of The 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death. IE'13*, 2013, https://doi.org/10.1145/2513002.2513006

Bendis, J. E., 'Developing educational virtual worlds with game engines', *ACM SIGGRAPH 2007 Educators Program on - SIGGRAPH '07*, 2007, https://doi.org/10.1145/1282040.1282068

Bongard-Blanchy, K., Rossi, A., Rivas, S., Doublet, S., Koenig, V., and Lenzini, G., '"I am definitely manipulated, even when I am aware of it. It's ridiculous!"—Dark patterns from the end-user perspective', *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, 2021, pp. 763–776 https://doi.org/10.1145/3461778.3462086

Brooks, G. A., and Clark, L., 'The gamblers of the future? Migration from loot boxes to gambling in a longitudinal study of young adults', *Computers in Human Behavior*, Vol. 141, 107605, 2023, https://doi.org/10.1016/j.chb.2022.107605

Castronova, E., *Exodus to the Virtual World: How Online Fun Is Changing Reality*, Palgrave Macmillan, 2007

Cave, E. M., 'What's wrong with motive manipulation?' *Ethical Theory and Moral Practice*, Vol. 10, Issue, 2, 2006, pp. 129–144, https://doi.org/10.1007/S10677-006-9052-4

D'Anastasio, C., 'How "Roblox" became a playground for virtual fascists', *WIRED*, 2021, https://www.wired.com/story/roblox-online-games-irl-fascism-roman-empire/

Cowan, G., Resendez, M., Marshall, E., & Quist, R. (2002). Hate Speech and Constitutional Protection: Priming Values of Equality and Freedom. Journal of Social Issues, 58(2), 247–263. https://doi.org/10.1111/1540-4560.00259

Dionisio, J. D. N., Burns, W. G., and Gilbert, R., '3D Virtual worlds and the metaverse: Current status and future possibilities', *ACM Computing Surveys* (CSUR), Vol. 45, Issue 3, 2013, https://doi.org/10.1145/2480741.2480751

Flanagan, M., Howe, D. C., and Nissenbaum, H., 'Values at play: Design tradeoffs in socially-oriented game design', *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2005, https://doi.org/10.1145/1054972

Friedman, B., 'Value-sensitive design', *Interactions,* Vol., 3, Issue 6, 1996, pp. 16–23. https://doi.org/10.1145/242485.242493

Gray, C. M., Chen, J., Chivukula, S. S., and Qu, L., 'End user accounts of dark patterns as felt manipulation', *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 2021, https://doi.org/10.1145/3479516

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., and Toombs, A. L., 'The dark (patterns) side of UX design', *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* - CHI '18, 2018, pp. 1–14, https://doi.org/10.1145/3173574.3174108

Hanna, J., 'Libertarian Paternalism, manipulation, and the shaping of preferences', *Social Theory and Practice*, Vol., 41, Issue 4, 2015, pp. 618–643

Jargon, J., 'Roblox struggles with sexual content. It hopes a ratings system will address the problem', *The Wall Street Journal*, 2021, https://www.wsj.com/articles/roblox-struggles-with-sexual-content-it-hopes-a-ratings-system-will-address-the-problem-11618660801

Kou, Y., and Gui, X., 'Flag and flaggability in automated moderation: The case of reporting toxic behavior in an online game community', *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–12, https://doi.org/10.1145/3411764.3445279

Kou, Y., and Gui, X, 'Harmful design in the Metaverse and how to mitigate it: A case study of user-generated virtual worlds on Roblox', *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 2023, pp. 175–188, https://doi.org/10.1145/3563657.3595960

Kou, Y., Hernandez, R. H. (Lindy), and Gui, X., '"The system is made to inherently push child gambling in my opinion": Child safety, monetization, and moderation on Roblox', *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–18, https://doi.org/10.1145/3706598.3713170

Kou, Y., Zhou, Y., Zhang, Z., and Gui, X., 'The ecology of harmful design: Risk and safety of game making on a metaverse platform', *Proceedings of the ACM Conference on Designing Interactive Systems 2024* (DIS 2024), 2024, pp. 1842–1856, https://doi.org/10.1145/3643834.3660678

Leopold, J., Lambert, J. R., Ogunyomi, I. O., and Bell, M. P., 'The hashtag heard round the world: How #MeToo did what laws did not', *Equality, Diversity and Inclusion*, Vol., 40, Issue 4, pp. 461–476, https://doi.org/10.1108/EDI-04-2019-0129/FULL/XML

Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., and Narayanan, A., 'Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites', *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 2019, 32, https://doi.org/10.1145/3359183

Pothong, K., Livingstone, S., Colvert, A., and Pschetz, L., 'Applying children's rights to digital products: Exploring competing priorities in design', *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*, 2024, pp. 93–104, https://doi.org/10.1145/3628516.3655789

Rowland, D., ''Virtual world, real rights?': Human rights and the internet', in *Emerging Areas of Human Rights in the 21st Century*, Routledge, 2021

Stolterman, E., 'The nature of design practice and implications for Interaction Design Research', *International Journal of Design*, Vol. 2, Issue 1, 2008, pp. 55–65.

Wang, Y., Leon, P. G., Scott, K., Chen, X., Acquisti, A., and Cranor, L. F., 'Privacy nudges for social media: An exploratory Facebook study', *Proceedings of the 22nd International Conference on World Wide Web*, 2013, pp. 763–770, https://doi.org/10.1145/2487788.2488038

Wiederhold, B. K., 'Sexual harassment in the Metaverse', *Cyberpsychology, Behavior, and Social Networking*, Vol., 25, Issue 8, 2022, pp. 479–480, https://doi.org/10.1089/cyber.2022.29253.editorial

Zendle, D., Meyer, R., and Over, H., 'Adolescents and loot boxes: Links with problem gambling and motivations for purchase', *Royal Society Open Science*, Vol. 6, Issue 6, 190049, 2019, https://doi.org/10.1098/rsos.190049

Zhang, Z., Moradzadeh, S., Gui, X., and Kou, Y., 'Harmful design in user-generated games and its ethical and governance challenges: An investigation of design co-Ideation of game creators on Roblox' *Proceedings of ACM Human-Computer Interaction,* 8(CHI PLAY), 2024, pp. 311:1-311:31. https://doi.org/10.1145/3677076

3.4. The Ethical, Governance, and Moderation Aspects of Virtual Worlds

Yubo Kou

The Pennsylvania State University

Abstract

This short report examines ethics, governance, and moderation as interrelated facets shaping user wellbeing in virtual worlds. As these immersive environments grow in scale and influence, especially among children, they present novel ethical dilemmas -ranging from predatory monetization and harmful avatar behaviors to the exploitation of platform affordances for harassment or manipulation. Governance is situated within both historical and contemporary contexts, highlighting how platform values, regulatory frameworks, and commercial incentives intersect to shape rules and user experiences. Moderation is a multi-layered process involving not only content and behavior, but also the design of virtual experiences themselves. This short report emphasizes the need for a context-aware, community-informed, and value-driven approach to ensure safety and inclusion. The report argues that safeguarding user wellbeing requires more than reactive enforcement; it demands integrated sociotechnical systems that embed ethical design, adaptive governance, and proactive moderation. In doing so, it raises open questions and calls for sustained research, platform accountability, and thoughtful regulation to support flourishing digital lives in complex virtual worlds.

Highlights

- Emerging ethical concerns are arising amid the rapidly evolving landscape of virtual worlds.
- The governance of virtual worlds reflects both historical foundations and contemporary developments.
- Moderation mechanisms must be reimagined to address emerging safety threats.

Introduction

User wellbeing is closely related to the ethical, governance, and moderation aspects of virtual worlds. What virtual worlds delineate as right or wrong, how virtual worlds set boundaries for user-generated content and user behaviors, as well as what are efficient sociotechnical mechanisms to uphold values and boundaries are all important questions to answer and directly shape how end users experience and feel as they enter virtual worlds. For example, a child virtual world user may initially be excited to enjoy a newly created virtual world and explore its immersive landscape, but fall prey to unethical design patterns such as predatory monetization (e.g., entering a room full of colorful, jumpy pets, all requiring a small amount of virtual currency (Kou & Gui, 2023)), and struggle to understand the virtual world's governance and moderation (Kou, Ma, et al., 2024) (e.g., whether groomers or cyberbullies exist, and how the virtual world can protect me from those malicious users). As such, user wellbeing hinges on the ethical, governance, and moderation aspects of virtual worlds.

In the rest of this short report, I will illustrate what these three aspects entail and their impact on user wellbeing. I will end the chapter with an integrated analysis of these aspects in relation to user wellbeing in virtual worlds.

Ethics in Virtual Worlds

The issue of ethics in virtual worlds stems from the very fact that virtual worlds present to our contemporary society a brand-new social situation, which we still have not known enough regarding how we should behave and what moral principles we should follow. Ethical dilemmas arise in numerous aspects of virtual worlds as users explore new ways to express themselves and to interact with each other, especially when any social encounters between two or more people are mediated by avatars (Hill, 2013). While users use avatars to represent themselves in virtual worlds, they identify with and get attached to their avatars (Wolfendale, 2007), further blending and reconfiguring the ethical boundaries between their real world and virtual world lives.

Thus, users are not just having fun in virtual worlds. They navigate complex ethical decision-making scenarios as they operate their avatars in various social encounters. For example, many virtual worlds allow users to customize their avatars, and ethical concerns arise when users configure their avatars for unethical self-expressions. One pertinent example is how Nazi roleplay could be enacted in certain virtual world platforms, where child players are encouraged to dress their avatars with Nazi uniforms (Keach, 2018). This raises significant ethical concerns as to what is the right way to customize avatars in virtual worlds, and how and to what extent virtual world platforms should intervene.

Avatar-mediated behaviors in virtual worlds could be ethically questionable as well. Virtual worlds provide users with great anonymity and loosened behavioral standards (Suler, 2004), and it becomes more likely for users to engage in unethical behaviors at the expense of other users' wellbeing or experience. For instance, our recent research on Roblox (Kou et al., 2025) reports that adult users might disguise themselves as child user and seek to groom minors, and that some users reach out to child users with the intention of scamming the latter out of their virtual currency.

What's more, virtual world platforms have been criticized for their unethical business models that encourage compulsive behaviors from their users, causing them to stay longer and spend more (Livingstone & Pothong, 2021). Our recent research suggests how virtual world platforms like Roblox allow the existence of chance-based mechanisms, which are gambling-like and seek to trick players into increased purchasing behavior (Kou, Zhou, et al., 2024). At a deeper level, virtual platforms can be profit-driven and only account for the financial interests of their shareholders, leading to unethical ways of monetize their users' experience (Kou, Zhou, et al., 2024).

Taken together, virtual worlds constitute an emergent sociotechnical context in which ethics is subject to ongoing negotiation. As users engage with novel affordances for self-presentation and social interaction, prevailing moral boundaries are reconfigured, often in ambiguous or contested ways. Platforms, as key mediators of these environments, respond with varying degrees of deliberation and accountability, and their interventions—or lack thereof—play a critical role in shaping normative practices. It is therefore insufficient to presume the ethical neutrality or adequacy of platform governance; rather, these systems must be interrogated as active agents in the construction of ethical norms. This ethically complex and dynamic terrain necessitates systematic inquiry and the development of robust theoretical frameworks to guide ethical understanding and practice in virtual worlds.

Virtual Worlds and Platform Governance

Virtual world governance has been an important topic ever since the beginning of the Internet where virtual environments came into existence. In the early 1990s, owners and moderators of multi-user dungeons (MUDs) grappled with the challenges of governing their virtual communities. Even back then, they already observed various disruptive users or behaviors. For example, Bartle's influential typology of players (Bartle, 1996) described one player type as killer, covering some of the users who derive fun from being aggressive towards other users. When running Habitat in the 1980s, one of the first large-scale virtual worlds, Morningstar and Farmer experimented with various strategies to protect their users' wellbeing against disruptive behaviors such as theft and cheating, and concluded that it was perhaps the best to observe how players evolve and support them to self-govern (Morningstar & Farmer, 2008). Those early day cases of virtual world governance showcase how dynamic the issue of governance is as we must take into account users' different needs, goals, and experiences and the particular sociotechnical configurations of the virtual world platforms.

Fast forward, the notion of 'platform' (Plantin et al., 2018) has taken a more important place in academic discourses around digital safety and wellbeing, capturing how platform affordances both support and constrain users' communication and expression. This notion can be helpful in unpacking the complexities of virtual world platform today, as it locates platform and the companies behind it as an important, if not the most, component in governing virtual worlds. Specifically, platform governance captures "the layers of governance relationships structuring interactions between key parties in today's platform society, including platform companies, users, advertisers, governments, and other political actors" (Gorwa, 2019). Drawing from this lens, we can argue that the governance of virtual worlds goes beyond managing what happens within a closed virtual environment, and is layered, and involves various actors within and beyond the virtual space. And this lens is crucial for developing a comprehensive understanding of today's virtual world governance.

Virtual world governance hinges on the underlying values and logics of platforms upholding them. When platforms center on profit maximization and push aside user wellbeing, this value orientation will permeate virtual world governance. Roblox, again, is a pertinent case here (Kou et al., 2025). The platform runs a revenue-sharing model with its users and incentivizes its users to generate and monetize virtual worlds. Partly as a result, its users are incentivized to embed a host of risky design features to extend their user engagement and to persuade their users to pay. For example, they implement gambling-like mechanisms like roulette and loot box to attract and engage child users. They embed sexually implicit content in their virtual worlds to pique child users' curiosity. In these ways, both creators of virtual worlds and the platform benefit financially, but their users' wellbeing deteriorates.

Platforms make and evolve policies for protecting user wellbeing in their virtual worlds, but the policymaking process can benefit from better deliberation. Online governance research has long observed a gap between platform policies and users' actual experiences, where the former tend to be static and lack changes, but the latter tend to be dynamic and evolving (Suzor, 2010). Platform policies do not necessarily capture what constitutes as disruptive behaviors as well as emergent forms of disruptive behaviors (Kou & Gui, 2021; Kou & Nardi, 2014). As a result, policies and their related enforcement mechanisms could fall behind in protecting user wellbeing. For example, while policies can explicitly prohibit verbal harassment, other emergent forms of harassment, such as 'teabagging,' where users use avatars to simulate sexual acts upon others' avatars, can fall out of the radar of policymaking. The rate of new forms of harassment emerging in virtual worlds can far outpace that of policymaking to articulate a list of unacceptable user behaviors.

Platforms' governance of virtual worlds also interacts with external governance bodies such as regional, national, and international authorities. Platforms need to comply with policy and regulatory frameworks at multiple layers, and sometimes with country-specific differences. For example, platforms may discourage certain means of monetization due to specific countries' regulations (e.g., Roblox, 2024a, 2024b) but allow them in others. This country-specific approach to policy compliance is strategic and follows the profit maximization value orientation but appears questionable in treating user wellbeing as overly context-dependent. There should be some universal values to uphold across country contexts. One of such approaches is the rights-based approach (Pothong et al., 2024) which identifies several fundamental human rights that apply across the world. This can serve as a policy framework and potentially guide the design of specific platform policies that prioritize user wellbeing.

Moderating Virtual Worlds

Moderation denotes a set of governance mechanisms that can discourage harm while encourage cooperation among community members (Grimmelmann, 2015). Moderation can be understood as the actual implementation of platform governance to shape user behavior. Moderation typically involves a three-step process: the detection, adjudication, and containment of harm. For instance, when a user engages in harassment, the first step is to detect the harmful behavior—either through automated systems or user-generated reports. Next, the moderation system adjudicates the case, determining whether the reported action violates established community standards or platform policies. If the behavior is deemed unacceptable, the final step usually involves punitive measures against the offending user, such as warnings, suspensions, or bans.

Moderation in virtual worlds happens at multiple levels. At the content level, content moderation techniques can be employed to detect and punish text-based or audio-based disruptive languages such as hate speech and verbal harassment (e.g., blocking a slur in a chat). At the behavioral level, we consider behavioral moderation where avatar-mediated behaviors must be properly regulated. This is exponentially harder than content moderation if we consider the "teabagging" case, where users can invent a myriad of ways to mimic teabagging. At the design level, design moderation becomes a necessity (Kou & Gui, 2023). In user-generated virtual worlds where users are allowed to design everything at different levels, it is possible that risks to user wellbeing come from not a single element in a virtual world, but from the holistic experience users gained by interacting with a virtual world. For example, users may design a slavery-themed virtual world (Zhang et al., 2024), and users only get to know the theme when they have immersed themselves in the virtual world for a while and start to recognize patterns of slavery through their holistic experience that includes interactions, comprehensions, and reflections. In this regard, design moderation accounts for moderation that can address harm and risk within the design practices that lead to problematic outcomes, rather than focus on the design outcomes.

While the novel sociotechnical context of virtual worlds gives rise to a variety of modalities that should be moderated, virtual world moderation needs to equally evolve into a multi-modal one that addresses not only harmful audio or text, but also harm originating from a combination of modalities (e.g., harmful behaviors that abuse certain design weaknesses) in a particular context. Virtual worlds are context-rich, compared to popular online platforms such as Reddit, Facebook, and Wikipedia, that rely primarily on text-based moderation. Thus, understanding the context within which harm takes place matters but involves careful considerations to be integrated into moderation (see Caplan, 2018).

No one knows the context of harm better than real users with lived experiences and experiential knowledge. Thus, this short report calls for a community-based approach, where community members' expertise and insights can be properly leveraged in moderation. Community involvement can be said to be already existent in some forms, such as flagging. However, they can play a more important role in moderation. For example, our research (Kou, Ma, et al., 2024) shows that online communities can help punished users to understand their violations and subsequently seek behavioral improvement. However, this role is not designed for in most moderation systems today.

Concluding Remarks

This short report has examined how ethics, governance, and moderation are not separate silos but deeply interconnected dimensions that collectively shape the wellbeing of users in virtual worlds. Ethical concerns—ranging from predatory monetization and avatar misuse to the exploitation of anonymity—are often enabled or constrained by governance structures. These, in turn, are made real through moderation practices, which must operate across content, behavior, and design.

A central theme emerging from this discussion is the entanglement of values, affordances, and enforcement. Ethical design must be embedded not only in abstract principles but in the very affordances of the virtual environment. Governance must be responsive and reflexive, aligning platform values with user wellbeing over shareholder interest. And moderation must evolve from reactive rule enforcement to context-sensitive, community-based interventions.

To ensure that virtual worlds remain safe for users, especially vulnerable ones such as children, platforms must embrace an integrated and value-driven approach. Ethical principles should guide platform governance. Governance structures must support nuanced, adaptive moderation. And moderation must be reimagined to encompass not just what users do, but what platforms allow and encourage them to build.

Returning to the child user entering a virtual world: safeguarding their wellbeing is not a matter of plugging isolated gaps. It requires a coherent sociotechnical ecosystem—where ethics are anticipated, governance is participatory and accountable, and moderation is layered, creative, and embedded. Only then can virtual worlds truly support flourishing, safe, and meaningful digital lives.

References

Bartle, R. A., 'Hearts, clubs, diamonds, spades: Players who suit MUDs', *The Journal of Virtual Environments*, Vol. 1, Issue 1.

Caplan, R., 'Content or context moderation? [Report]. Data & Society Research Institute, 2018, https://apo.org.au/node/203666

Gorwa, R., 'What is platform governance?', *Information, Communication & Society*, Vol. 22, Issue 6, 2019, pp 854–871, https://doi.org/10.1080/1369118X.2019.1573914

Grimmelmann, J., 'The Virtues of Moderation', *Yale Journal of Law and Technology*, Vol. 17, 2015.

Hill, D. W., 'Avatar ethics: Beyond images and signs', *Journal for Cultural Research*, Vol. 17, Issue 1, 2012, pp. 69–84. https://doi.org/10.1080/14797585.2012.719689

Keach, S., 'Roblox kids' game is a haven for twisted Jihadi, Nazi and KKK racist roleplay' The Sun, 2018 https://www.thesun.co.uk/tech/6710158/roblox-game-racist-jihad-nazi-kkk-racism-twin-towers-children/

Kou, Y., and Gui, X., 'Flag and flaggability in automated moderation: The case of reporting toxic behavior in an online game community', *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems,* 2021,

Kou, Y., and Gui, X., 'Harmful design in the metaverse and how to mitigate it: A case study of user-generated virtual worlds on Roblox', *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 2023, pp. 175–188. https://doi.org/10.1145/3563657.3595960

Kou, Y., Hernandez, R. H. (Lindy), and Gui, X., '"The system is made to inherently push child gambling in my opinion": Child safety, monetization, and moderation on Roblox', *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp 1–18. https://doi.org/10.1145/3706598.3713170

Kou, Y., Ma, R., Zhang, Z., Zhou, Y., and Gui, X., 'Community begins where moderation ends: Peer support and Its implications for community-based rehabilitation', *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–18. https://doi.org/10.1145/3613904.3642675

Kou, Y., and Nardi, B., 'Governance in League of Legends: A hybrid system', Foundations of Digital Games, 2014.

Kou, Y., Zhou, Y., Zhang, Z., and Gui, X., 'The ecology of harmful design: Risk and safety of game making on a metaverse platform', *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, 2024, pp. 1842–1856. https://doi.org/10.1145/3643834.3660678

Livingstone, S., and Pothong, K., 'Playful by design: Free play in a digital world [Monograph].' Digital Futures Commission, 5Rights Foundation, 2021, https://digitalfuturescommission.org.uk/wp-content/uploads/2021/11/A-Vision-of-Free-Play-in-a-Digital-World.pdf

Morningstar, C., and Farmer, F. R.,. 'The lessons of Lucasfilm's Habitat', *Virtual Worlds Research: Past, Present and Future*, Vol. 1, Issue 1, 2008, https://doi.org/10.4101/jvwr.v1i1.287

Plantin, J.-C., Lagoze, C., Edwards, P. N., and Sandvig, C., 'Infrastructure studies meet platform studies in the age of Google and Facebook', *New Media & Society*, Vol. 20, Issue 1, 2018, pp. 293–310, https://doi.org/10.1177/1461444816661553

Pothong, K., Livingstone, S., Colvert, A., and Pschetz, L., 'Applying children's rights to digital products: Exploring competing priorities in design', *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*, 2024, pp. 93–104, https://doi.org/10.1145/3628516.3655789

Roblox, 'Update on paid random items restriction for UK users under 18', Developer Forum | Roblox, 2024a, https://devforum.roblox.com/t/update-on-paid-random-items-restriction-for-uk-users-under-18/3072183

Roblox, 'Update on paid random items restriction for Australian users', Developer Forum | Roblox, 2024b, https://devforum.roblox.com/t/update-on-paid-random-items-restriction-for-australian-users/3153602

Suler, J., 'The online disinhibition effect', *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, Vol. 7, Issue 3, 2004, pp 321–326, https://doi.org/10.1089/1094931041291295

Suzor, N. P., 'The role of the rule of law in virtual communities', *Berkeley Technology Law Journal*, Vol. 25, Issue 4, 2010, pp 1818–1886.

Wolfendale, J., 'My avatar, my self: Virtual harm and attachment', *Ethics and Information Technology*, Vol. 9, Issue 2, 2007, pp. 111–119. https://doi.org/10.1007/s10676-006-9125-z

Zhang, Z., Moradzadeh, S., Gui, X., and Kou, Y., 'Harmful design in user-generated games and its ethical and governance challenges: An investigation of design co-ideation of game creators on Roblox', Proceedings of ACM Human-Computer Interaction, 8, 2024, pp 311:1-311:31, https://doi.org/10.1145/3677076

4. Towards Methodological Innovation

4.1. Vanishing Effects? On the Importance of Assessing the Effects of Virtual Worlds on Users' Well-being among Early Adopters and First-time Users[7]

Susanne E. Baumgartner

Amsterdam School of Communication Research, University of Amsterdam

Abstract

Virtual worlds are likely to have a profound impact on individuals' well-being. Although existing research on the effects of digital media on well-being might be informative to guide future research on digital virtual worlds, it is important to acknowledge shortcomings of previous theoretical as well as empirical considerations on the effects of digital media. Particularly, I will argue in this short report that previous research has likely underestimated the effects of digital media on well-being due to a limited empirical conceptualization of media effects that primarily considered media effects as linear. Based on established psychological theories this assumption is, however, highly unlikely. Specifically, based on theories on habituation and adaptation, it is likely to expect that the effects of virtual worlds on well-being stabilize after some time. In this short report, I will delineate how habituation and adaption processes are crucial in understanding the effects of virtual worlds on well-being. Importantly, the proposed view has also strong consequences for future studies on virtual worlds. I will argue that to truly understand the effects of virtual worlds on well-being we need a new generation of studies that examines effects among first-time users of such technologies. This will allow us to capture the true effects of virtual worlds before effects stabilize and become elusive.

Highlights

- Effects of virtual worlds on well-being likely vanish over time
- Effects of virtual worlds on well-being are only detectable during effect-sensitive periods
- Studies need to assess effects on well-being among first-time users or early adopters

With big tech companies heavily investing in the development of virtual worlds, it is likely that they will significantly shape our future media landscape. As virtual worlds have the potential to be mass-adopted and to considerably disrupt the current media landscape, it is vital to understand the potential consequences of these technologies for their users' well-being. Future research can thereby highly profit from existing knowledge that the media effects field has garnered in the past

---

[7] This short report is based on ideas presented in this paper: Baumgartner, S. E. (submitted for publication). Why we see media effects but do not find them: Media Effects Stabilize After Repeated Media Exposure. *Communication Theory*.

decades. This provides the perfect opportunity to direct future research into the effects of virtual worlds, and set-up the most efficient and targeted research projects. Importantly, these insights might help to investigate the potential effects of virtual worlds before they are mass-adopted, and might thus help to guide policies to prevent negative effects.

The potential positive as well as negative effects of media on well-being have been studied for many decades and for various types of media, such as TV (Cantor & Mares, 2001), video games (Halbrook et al., 2019), social media (Dienlin & Johannes, 2022), and mobile media (Schneider et al., 2022). Although the public typically voices strong concerns about the potential detrimental effects of digital media, the research field oftentimes lacks clear conclusions on the size of such effects. For example, although there are strong concerns about the detrimental impact of social media on youth' mental health, recent meta-analyses and reviews of the literature conclude that the effects are "very small" and "weak" (e.g., Appel et al., 2020; Dienlin & Johannes, 2022). The phenomenon of "minimal effects" is not new but has been shown repeatedly throughout the history of the field (e.g., Lang & Ewoldson, 2009). Although several explanations have been put forward for why studies typically find only minimal effects (e.g., differences among persons, measurement issues), I argue that an important reason is that previous research mainly studied linear effects, and ignored how effects develop over time.

The vast majority of studies on the effects of digital media on well-being are based on correlational designs that test whether the frequency of engagement with digital media is linearly related to a specific outcome. For example, it is tested whether individuals who currently use social media more frequently have lower levels of well-being than individuals who use social media less frequently (Orben et al., 2019). These research designs are typically based on survey studies, and have specific advantages; they are relatively cheap and can assess these relationships among large groups of people. However, these designs are also problematic for various reasons.

Most importantly, I argue that the linearity assumption (i.e., the more someone uses a specific technology, the stronger the effects) inherent to these designs is highly unlikely. Both, theoretical conceptualizations in the media effects literature (e.g., Slater et al., 2015; Shehata et al., 2021; Shehata et al., 2024) as well as well-established psychological theories on the effects of well-being (Luhmann & Intelisano, 2018; Rankin et al., 2009), propose that effects of digital media on well-being stabilize after some time. This means that effects on well-being do emerge at the beginning and then stabilize or even fade over time. Considering this stabilization is crucial because it has strong consequences for how we need to set-up future studies.

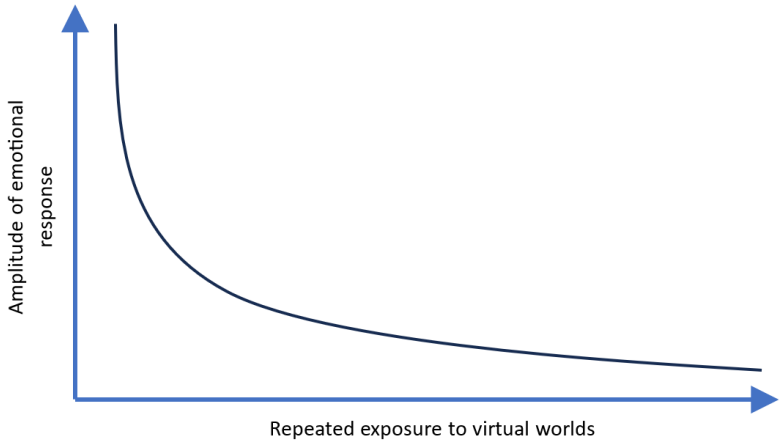
Stabilizing effects on well-being: Habituation and adaption to media effects

Based on well-established psychological theories it is likely that digital media (including virtual worlds) might have initial negative effects on well-being. However, theories on 1) habituation (Rankin et al., 2009), and 2) adaptation (Luhmann & Intelisano, 2018) propose that these effects should stabilize after some time.

Habituation to repeated stimuli is one of the most robust and consistent effects reported in the psychological literature (see for example: Grill-Spector et al., 2006). Habituation is defined as a decrease in physiological, psychological, and behavioral responses to repeated stimuli (Figure 1; e.g., Rankin et al., 2009). This means that if individuals are repeatedly exposed to the same or similar stimuli, their responses to this stimulus decline over time. This has been shown for a variety of stimuli including emotion-evoking (e.g., Ferrari et al., 2020; Grissom & Bhatnagar, 2009) or pain-evoking stimuli (Rennefeld et al., 2010). In the media effects field, this has been shown for example

for the effects to violent videos. For instance, Grizzard et al. (2015) studied the effects of repeated exposure to violent video games. In their study, participants played the same violent game for 10 minutes on four days in a row. They showed that participants who played a violent video game for the first time showed increased arousal levels on the first day but these arousal levels declined over the following three days. Importantly, Grizzard et al. (2015) also reported a generalizability of the effect to another violent video game played on the 5th day. Participants showed habituated physiological responses to this new but similar violent game as well. This study, thus, shows that initial reactions to media stimuli decline over time. This indicates that individuals -when exposed frequently to similar media stimuli– might react less strong over time. The strength of the impact of media stimuli thus weakens over time.

Figure 1. Habituation of emotional and physiological responses



Note. The figure depicts a prototypical media effect trajectories after repeated exposure based on habituation. The initial emotional response decreases with additional exposures, and stabilizes at low levels.

Source: Author's own elaboration

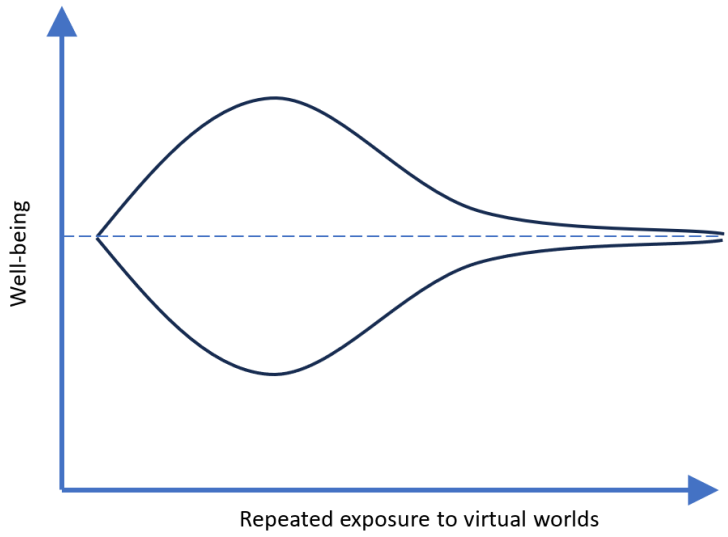
While habituation explains fading media effects on the physiological and emotional level, hedonic adaptation processes might be particularly useful in understanding effects on well-being (see Figure 2). Hedonic adaptation is a well-established phenomenon that shows that people quickly adapt to repeated negative as well as positive experiences (Frederick & Loewenstein, 1999). For example, it has been shown that although negative or positive life events can change an individuals' level of life satisfaction dramatically, individuals tend to return to their initial happiness levels after some time (e.g., Lucas, 2007).

Hedonic adaptation processes are thus crucial for our understanding of the potential effects of virtual worlds on well-being. Considering the key assumptions of hedonic adaptation –a return of life satisfaction to initial levels after negative or positive life events– it is also likely to assume that digital media also have only fleeting effects on well-being. Even if they negatively affect well-being, after some time, individuals might adapt to these effects. If virtual worlds lead to a decrease in well-being, it is likely that individuals over time cognitively and behavioral adapt to these circumstances. This means that over time, even if virtual worlds initially decreased well-being, individuals might adapt and thus might not experience the effects as strongly anymore. At that

point, the effects on an individual's well-being might not be detectable anymore, even if there was a strong effect previously.

Adaptation to digital media might be explained by automatic, cognitive processes (e.g., Lucas, 2007) but also by behavioral adaptations. For example, the mass adoption of smartphones has led to a variety of behavioral changes to incorporate these devices more seemingly into our daily routines. For instance, many individuals manage their notifications in a way that they are not constantly interrupted by incoming messages (e.g., turning off the sound). These adaptive behaviors might prevent or diminish potential negative responses to media. Implementing these adaptation strategies might lead to paradoxical observations: We might witness a dramatic change in our daily behaviors due to digital media but at the same time these behavioral changes are a reflection of adaptation processes that prevent us from experiencing long-lasting effects on our well-being. It is likely to assume that similar adaptation processes will take place with the mass adoption of virtual worlds.

Figure 2. Adaptation to virtual worlds



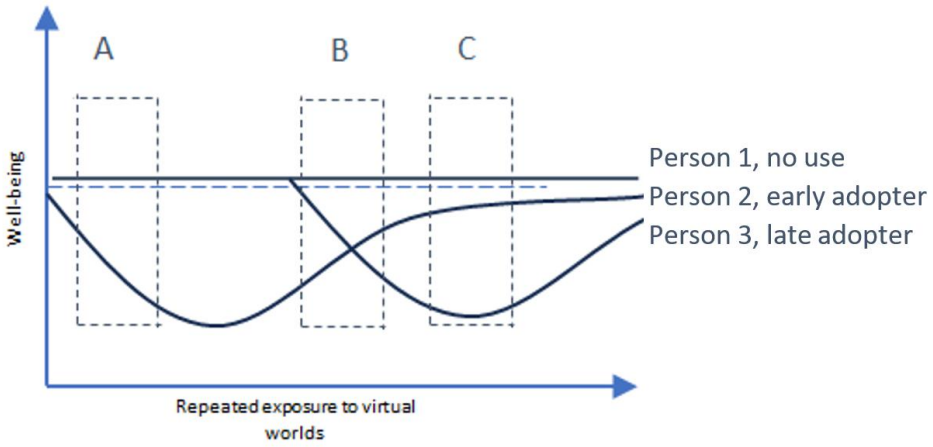
Note. The figure depicts prototypical media effect trajectories after repeated exposure based on hedonic adaptation. Negative as well as positive responses decrease over time and then stabilize.

Source: Author's own elaboration

Based on habituation and adaptation, it is likely that effects of virtual worlds on well-being stabilize at some point. This stabilization thus means that effects are visible only during specific effect-sensitive periods (see Figure 3). Consider the three prototypical persons in Figure 3; Person 1 has never used virtual worlds and has stable levels of well-being. Person 2 has started using virtual worlds several months ago. Although that person felt initially less happy, the person adapted to their use of virtual worlds and their levels of well-being increased again. Person 3, just started using virtual worlds a few weeks ago, their level of well-being decreased and that person had no time yet to adapt to these changes. Importantly, if we conduct a study during the time period depicted in Box C, we won't find any within-person effects of virtual worlds on well-being as for all three persons, the effects are currently relatively stable. However, if we conduct the same study, during the effect-

sensitive time-periods depicted in Box A or B, when Persons 1 and 2 started using virtual worlds, we will be able to detect effects for these persons.

Figure 3. Stabilization of Effects and its Consequences for the Timing of Studies



Note. The figure depicts the effect trajectories of three prototypical persons who start using virtual worlds at different points in time. Person 1 does not use virtual worlds, and has stable levels of well-being. Person 2 was an early adopter. Well-being of Person 2 decreased and at some point increased again. Person 3 started using virtual worlds later on. During period C, well-being for all three individuals is stable on varying levels. The negative effects of using virtual worlds on well-being are only detectable during Time A for Person 2, and during Time B for Person 3.

Source: Author's own elaboration

Thus, the main conclusion that we can draw from this theorizing is that the timing of assessment of potential effects is crucial. We need to conduct studies during effect-sensitive periods to being able to detect effects for individual persons. These effect-sensitive periods are likely to occur when individuals are exposed to technologies for the very first time. Relatedly, studies ideally should assess effects over longer periods of time to not only assess initial effects but also to understand how these effects develop over the course of time, and whether and how adaptation processes set in.

The importance of studying effects among early adopters and first-time users

The idea of stabilizing media effects has strong implications for future attempts to study the effects of virtual worlds. To capture the effect-sensitive period and the following effect patterns, it is necessary to conduct longitudinal studies among individuals who start using virtual worlds for the first time. These types of studies can be either conducted in the real environments of people, or in controlled environments.

In **controlled settings**, this would mean to invite participants with limited prior experience of virtual worlds into the lab and let them use these technologies for a longer period of time. The studies conducted for the effects of action video games on cognitive skills could be considered as ideal examples of how to do this (e.g., Bediou et al., 2018). Participants in these studies are asked to

play a game for a given number of hours (e.g., 60 hours for 2-4 weeks). Participants' cognitive skills are measured and compared before and after the training period. These studies are ideal because they do control for individuals' prior experience and exclude participants with gaming histories. However, from an ethical perspective these types of studies are not desirable if one is interested in potential negative effects on well-being. For example, when interested in the effects of virtual worlds on loneliness or depression, we can ethically not expose individuals to these technologies in the context of a study.

An alternative option would be to conduct **field studies** with groups of individuals who start using virtual worlds in their natural environments. For example, it can be incredible informative to study early adopters of a technology (e.g., Kraut et al, 1998). These early adopters use these technologies in an environment that has not yet mass-adopted these technologies. Therefore they offer unique insights into how these technologies are used, how individuals are affected, and how individuals adapt to potential adverse effects. These studies are ideal because they can inform policies even before these technologies are mass-adopted and potentially impact large user groups. A few studies have been conducted among early internet and TV users (e.g., Kraut et al., 1998; Williams, 1986), however, most research efforts have set in much later when technologies were already used by large parts of society.

Once technologies are mass-adopted, studies among early adopters are not possible anymore. In that case, it is crucial to study first-time users, that is, individuals who start using a technology for the first time. For example, studies could examine children, teenagers or young adults who start using specific technologies for the first time. This can be studied for example among families who receive virtual devices for the first time (see for example Weis & Cerankosky, 2010).

For these types of field studies, it is necessary to recruit individuals who start using such technologies and follow them for a longer period of time (i.e., several months/years). Such studies do not only allow to study initial effects, but also to trace adaptation processes as they occur in real life. These studies require tools to objectively trace individuals' use of and behavior within the virtual worlds, as well as subjective indicators of their well-being (for example, with diary entries, or short surveys several times during the study period). These types of studies can be considered ideal but require substantial research effort for the recruitment and selection of participants, as well as for keeping participants compliant for longer periods of time.

Conclusion

Well-established media effects theories as well as psychological theories predict that the effects of exposure to virtual worlds likely stabilizes at some point. This has strong consequences for empirical endeavors to detect these effects because effects can likely solely be detected during effect-sensitive periods before these effects stabilize. To truly understand the effects of virtual worlds on well-being we, thus, need a new generation of studies that examine effects among first-time users of emerging technologies. This will allow us to capture the true effects of virtual worlds before effects stabilize and become elusive. This idea of vanishing media effects is not new but has been theorized several decades ago (Gerbner et al., 1986). Nevertheless, studies to capture the initial effects of digital media have been rare. This is problematic because if new technologies are mass-adopted, as has happened for example with television, the internet, and smartphones, it is likely that over time, effects become more difficult to detect as everyone might be affected, and at some point, the effect pattern for each individual should stabilize. This is in line with Morgan's (1986) conclusion that the longer we live with a medium, "the smaller its observable impact may become"

(p. 135). It is thus crucial to detect the effects of emerging technologies early on, so that potential negative effects can be detected and policies can effectively target detrimental effect before these technologies are mass-adopted.

References

Appel, M., Marker, C., and Gnambs, T. 'Are social media ruining our lives? A review of meta-analytic evidence', *Review of General Psychology*, Vol. 24, Issue 1, 2020, pp. 60-74, https://doi.org/10.1177/1089268019880891

Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S., and Bavelier, D. 'Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills', *Psychological Bulletin*, Vol. 144, Issue 1, 2018, pp. 77–110, https://doi.org/10.1037/bul0000130

Cantor, J., and Mares, M. L., 'Effects of television on child and family emotional well-being', in: Jennings, B. and Alison, B. J. (eds), *Television and the American family*, Lawrence Erlbaum Associates, 2001, pp. 317-332.

Dienlin, T., and Johannes, N., 'The impact of digital technology use on adolescent well-being', *Dialogues in Clinical Neuroscience*, Vol. 22, Issue 2, 2022, pp. 135–142, https://doi.org/10.31887/DCNS.2020.22.2/tdienlin

Ferrari, V., Mastria, S., and Codispoti, M., 'The interplay between attention and long-term memory in affective habituation', *Psychophysiology*, Vol. 57, Issue 6, 2020, e13572, https://doi.org/10.1111/psyp.13572

Frederick, S., and Loewenstein, G., 'Hedonic adaptation', in: Kahneman, D., Diener, E., and Schwarz, N. (eds*), Well-being: The foundations of hedonic psychology*, Russell Sage Foundation, 1999, pp. 302–329.

Gerbner, G., Gross, L., Morgan, M., and Signorielli, N., 'Living with television: The dynamics of the cultivation process', in: Bryant, J., and Zillman, D. (eds), *Perspectives on media effects*, Erlbaum, 1986, pp. 17-40.

Grill-Spector, K., Henson, R., and Martin, A., 'Repetition and the brain: neural models of stimulus-specific effects', *Trends in Cognitive Sciences*, Vol. 10, Issue 1, 2006, pp. 14-23, https://doi.org/10.1016/j.tics.2005.11.006

Grissom, N., and Bhatnagar, S., 'Habituation to repeated stress: get used to it', *Neurobiology of Learning and Memory*, Vol. 92, Issue 2, 2009, pp. 215-224, https://doi.org/10.1016/j.nlm.2008.07.001

Grizzard, M., Tamborini, R., Sherry, J. L., Weber, R., Prabhu, S., Hahn, L., and Idzik, P., 'The thrill is gone, but you might not know: Habituation and generalization of biophysiological and self-reported arousal responses to video games', *Communication Monographs*, Vol. 82, Issue 1, 2015, pp. 64-87, https://doi.org/10.1080/03637751.2014.971418

Halbrook, Y. J., O'Donnell, A. T., and Msetfi, R. M., 'When and how video games can be good: A review of the positive effects of video games on well-being', *Perspectives on Psychological Science*, Vol. 14, Issue 6, 2019, pp. 1096-1104, https://doi.org/10.1177/174569161986380

Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukophadhyay, T., and Scherlis, W., 'Internet paradox: A social technology that reduces social involvement and psychological well-being?', *American Psychologist*, Vol. 53, Issue 9, 1998, pp. 1017, https://doi.org/10.1037/0003-066X.53.9.1017

Lang, A., and Ewoldsen, D., 'Beyond effects: Conceptualizing communication as dynamics, complex, nonlinear, and fundamental', in: Allen, S. (Ed.), *Rethinking Communication Keywords in Communication Research*, Hampton Press, New York, 2009, pp. 109–120.

Lucas, R. E., 'Adaptation and the set-point model of subjective well-being: Does happiness change after major life events?' *Current Directions in Psychological Science*, Vol. 16, Issue 2, 2007, pp. 75-79. https://doi.org/10.1111/j.1467-8721.2007.0047

Luhmann, M., and Intelisano, S., 'Hedonic adaptation and the set point for subjective well-being', in: Diener, E., Oishi, S., and Tay, L. (eds), *Handbook of Well-Being.* Noba Scholar Handbook series: Subjective well-being, DEF Publishers, Salt Lake City, 2018, pp. 195-219.

Morgan, M., 'Television and the erosion of regional diversity', *Journal of Broadcasting & Electronic Media*, Vol. 30, Issue 2, 1986, pp. 123-139. https://doi.org/ 10.1080/08838158609386615

Orben, A., Dienlin, T., & Przybylski, A. K., 'Social media's enduring effect on adolescent life satisfaction', *Proceedings of the National Academy of Sciences*, Vol. 116, Issue 21, 2019, pp. 10226-10228, https://doi.org/10.1073/pnas.1902058116

Rankin, C. H., Abrams, T., Barry, R. J., Bhatnagar, S., Clayton, D. F., Colombo, J., ... and Thompson, R. F., 'Habituation revisited: an updated and revised description of the behavioral characteristics of habituation', *Neurobiology of Learning and Memory*, Vol. 92, Issue2, 2009, pp. 135-138, https://doi.org/10.1016/j.nlm.2008.09.012

Rennefeld, C., Wiech, K., Schoell, E. D., Lorenz, J., & Bingel, U., 'Habituation to pain: further support for a central component', Pain, Vol. 148, Issue 3, 2010, pp. 503-508, https://doi.org/10.1016/j.pain.2009.12.014

Schneider, F. M., Lutz, S., Halfmann, A., Meier, A., and Reinecke, L., 'How and when do mobile media demands impact well-being? Explicating the integrative model of mobile media use and need experiences (IM3UNE)', *Mobile Media & Communication*, Vol. 10, Issue 2, 2022, pp. 251-271, https://doi.org/10.1177/205015792110549

Shehata, A., Andersson, D., Glogger, I., Hopmann, D. N., Andersen, K., Kruikemeier, S., & Johansson, J., 'Conceptualizing long-term media effects on societal beliefs', *Annals of the International Communication Association*, Vol. 45, Issue 1, 2021, pp. 75-93, https://doi.org/10.1080/23808985.2021.1921610

Shehata, A., Thomas, F., Glogger, I., & Andersen, K., 'Belief maintenance as a media effect: a conceptualization and empirical approach', *Human Communication Research*, Vol. 50, Issue 1, 2024, pp. 1-13, https://doi.org/10.1093/hcr/hqad033

Slater, M. D., 'Reinforcing spirals model: Conceptualizing the relationship between media content exposure and the development and maintenance of attitudes', *Media Psychology*, Vol. 18, Issue 3, 2015, pp. 370-395, https://doi.org/10.1080/15213269.2014.897236

Weis, R., & Cerankosky, B. C., 'Effects of video-game ownership on young boys' academic and behavioral functioning: A randomized, controlled study', *Psychological Science*, Vol. 21, Issue 4, 2010, pp. 463-470, https://doi.org/10.1177/0956797610362670

Williams, T. M., *The impact of television: A natural experiment in three communities*, Academic Press, 1986.

4.2. Credible and Transparent Industry-academia Collaborations for Understanding Life Online

Matti Vuorre

Department of Social Psychology, Tilburg University

Abstract

Digital devices, platforms, and media have become adjacent, if not central, to most domains of modern life. Concurrent to these technologies' global penetration, concerns about their effects on users' psychological functioning have also become commonplace, yet the relations remain poorly understood. Because of shortcomings in current methodologies and resulting lack of confidence in the existing evidence base, many investigators have converged on a central feature of online platforms as a promising path towards a better understanding of lives online: The unprecedented amount of information about users' behavior collected by internet-powered platforms and devices. While these digital footprints data are typically collected for development and monetization purposes, scientists increasingly recognize their potential for scientific study of human behavior for the common good. While several inroads for independent scientists' access to industry data have been laid and pilot studies conducted, technical, practical, legal, and ethical issues must be solved before the promises of these data for transparent and credible industry-academia collaborations can be fully realized. I call for a collective effort to overcome these obstacles to enable the responsible use of people's digital footprints data for understanding and potentially bettering human experiences both on- and offline.

Highlights

- Prevalence of online technologies has inspired widespread concerns over their effects on users' psychological functioning
- Results from current efforts on technologies impacts on their users remain mixed and conclusions are uncertain at best
- Researchers increasingly recognize the potential value of digital footprints data—the automatically collected telemetry on user behaviors
- Initial efforts to use digital footprints data to better understand life online show promise, but much remains to be done

Introduction

According to the International Telecommunication Union's estimates, 5.5 billion individuals, or 68% of the global population, were internet users in 2024 (International Telecommunications Union, 2024). Compared to fewer than a billion users two decades earlier, this dramatic uptake, along with the concurrent evolution of internet-enabled devices, services, and platforms (e.g., smartphones, online banking, and social media, respectively) has affected most domains of human life and

blurred the distinction between offline and online: Already in 2018 half of US teenagers reported being online "almost constantly" (Anderson and Jiang, 2018). In turn, this global digitalization of human behavior has prompted widespread concern and speculation about these omnipresent technologies' effects on human behaviors, cognitions, social lives, and well-being.

Because these technologies, and concerns over their negative effects, are so widespread, it is critical to collect and disseminate the best possible evidence regarding their roles in human well-being and functioning. However, in contrast to salient popular writings that suggest conclusive evidence for negative effects, current scientific evidence is uncertain and based on a largely cross-sectional and unrepresentative body of research that cannot support causal conclusions of digital technologies effects' on individuals' well-being (Orben, 2020; Odgers, 2024). Moreover, the vast majority of studies have ignored the Global South where the uptake of these technologies is currently most rapid, thereby limiting our understanding of the issues in their most important contexts (Ghai et al., 2022; Ghai et al., 2023).

These worries, fueled by preliminary reports of digital ills that were later found methodologically wanting (Appel, Marker, and Gnambs, 2020; Ophir, Lipshits-Braziler, and Rosenberg, 2020; Twenge et al., 2017) and highly variable between individuals (Valkenburg et al., 2021), have largely not been substantiated by subsequent empirical scientific inquiry (Hancock et al., 2022; Valkenburg, Meier, and Beyens, 2022). Instead, these efforts have pointed toward small statistical correlations that cannot be directly interpreted as meaning that online technologies cause psychological ills, and great uncertainty because of the relatively poor quality of existing studies (Appel, Marker, and Gnambs, 2020; Hancock et al., 2022; Best, Manktelow, and Taylor, 2014; Dickson et al., 2019; Orben, 2020; National Academies of Sciences, 2024).

Nevertheless, this scientific uncertainty and relative lack of empirical support for evidence of causal effects has not prevented policymakers from acting: Many countries have adopted health advisories and guidelines on young people's "screen time" (Kaye et al., 2020), with China and South Korea, among others, outright restricting adolescents' digital activities (The Chosun Daily, 2024; BBC News, 2021). Most recently, Australia moved towards banning under-16s from social media platforms (Wilson, 2024). Facing mixed evidence and pressures to inform evidence-based policymaking, many scientists have recognized the potential value of one of the most criticized features of online technologies: The collection and processing of large-scale and detailed data on user behaviors—digital footprint data. Instead of using those data for commercial purposes—typically the reason that the data are collected—researchers have reasoned that these data can alleviate many methodological shortcomings of previous studies on digital technologies' psychological effects (Johannes, Vuorre, and Przybylski, 2021).

Studies of online activities in relation to psychological well-being usually rely on self-reports of online behavior. Typically, study participants report their subjective perceptions of their time spent online (or on social media, playing video games, on smartphones, etc.) in a given time window, which typically spans from the past day to the past year (Parry et al., 2021). Alternatively, several questionnaire scales exist that attempt to probe these experiences beyond time spent, often with a focus on negative experiences such as problematic use patterns (Ellis, 2019). These self-reported measures of online behaviors are then statistically correlated with self-reported psychological outcomes ranging from affect and loneliness to depression and self-harm.

First, this broad methodology suffers from a general shortcoming of survey research: Common method bias, which can artificially make associations between variables appear greater simply because they are reported on using the same instrument (e.g. ticking boxes in an online form; Podsakoff et al., 2024). Second, besides general measurement-related issues such as reliability and

validity, self-reports of digital media use, specifically, are known to be noisy (i.e. they correlate only moderately with objectively recorded behavior; Parry et al., 2021; Ellis et al., 2019) and biased (e.g. more depressed individuals are less accurate in estimating their time spent on social media; Sewall et al., 2020; Sewall and Parry, 2021). These systematic errors in self-reported digital technology use measures has prompted many researchers to look beyond self-reports and toward objective records of online behaviors.

Digital footprints

What are digital footprints?

As modern digital technologies offer an increasingly varied menu of affordances to their users, they also record and analyze nearly all the users' behaviors. For example, banks have detailed records of their customers' electronic purchase histories; social media platforms record for how long users attend to content and whether/how they interact with it; video game platforms record detailed information on what players are doing and when. These data are typically collected for commercial (e.g. advertisement) and development (e.g. product innovation) purposes, but their societal and scientific value are increasingly recognized, and many scientific efforts have produced valuable information about people's online behaviors using those data through various means.

Existing efforts to better understand lives online using digital footprints data

Instagram Data Access Pilot for Well-being Research

Social media platforms, such as Meta's Instagram, feature heavily on discussions about internet technologies' potential psychological effects. A promising ongoing project targeting this concern is a collaboration between the academic nonprofit Center for Open Science, Meta, and an independent editorial board (on which I serve). In this Instagram Data Access Pilot for Well-being Research (https://www.cos.io/meta) teams of academic scientists will independently recruit study participants, and then request and receive those users' detailed Instagram use data from a "menu" of potential data communicated by Meta (Meta Platforms, 2024).

The management of these projects is based on the Registered Report format, whereby research teams submit their project for evaluation to the editorial board, who then recruits reviewers to evaluate the work before any data has been collected. If a project passes this stage, authors submit their data request to the Center for Open Science, which then coordinates the data donation with Meta—thus ensuring the independence and anonymity of the scientific team in executing their research to scientific ideals that are not communicated nor thereby influenced by the industry stakeholder. After this process, the team can investigate their sample's detailed Instagram use footprints alongside any psychological information they choose to collect, thus leading to valuable and detailed studies of objective social media behaviors in relation to psychological outcomes.

Data donation studies (datadonation.eu)

Many local regulations, most prominently the GDPR in the EU, require large online platforms to allow users to request data the platform has collected about their behavior. As a response, many platforms have made exporting such Data Download Packages (DDPs) easy for their users. These DDPs have then been adopted as an object of scientific inquiry in data donation studies, where researchers recruit participants to obtain their DDPs from specific platforms and donate them to the

researchers. For example, datadonation.eu provides free and open source resources and software for researchers to automate privacy-friendly processing of such data on users' own computers, which can then be readily analyzed to better understand the participants' online behaviors. The datadonation.eu project has already supported promising empirical projects, such as examining parents' motivations to use virtual assistants when interacting with their children (Wald et al., 2023).

Application Programming Interfaces (APIs)

Many online platforms serve publicly available Application Programming Interfaces (APIs) that facilitate data use within- and across platforms through the internet. For example, the popular video game platform Steam has an API that can be queried for e.g. information about how many players are currently playing specific games on the platform. These data are useful beyond the company's own interests, because third parties can use the data to e.g. present players with information on trending games that they might be interested in (e.g. steamdb.info). In one study, we collaborated with the developer of one of these third-party API clients (SteamDB) to investigate large scale play patterns during COVID-19 (Vuorre et al., 2021) and found increased multiplayer game engagement during the pandemic.

More generally, such APIs make possible seamless integration of data between services, and researchers can therefore link digital footprint data obtained through APIs with other data of scientific interest, such as psychological survey instruments. Many researchers are currently working on implementing online platforms for linking those data with digital footprints acquired through these API data sources, and I anticipate the results of these efforts to be illuminating with respect to the roles that those online technologies play in individuals' psychological well-being.

Academia-industry collaborations

In addition to more systematic efforts, many research projects have been successfully realized by direct collaboration between a team of academics and an industry stakeholder in possession of digital footprints data. Some examples include my work with video game industry partners: In one study, seven global video game publishers recruited players, via marketing emails, who then agreed for the publishers to make their gameplay data available to us and to complete our psychological surveys (Vuorre et al., 2022). In this manner, we were able to link objective gameplay records with participants' responses to psychological instruments and publish all data openly for other researchers to investigate. In another collaboration with the UK-based developer FuturLab, we developed and published a modified version of their commercially available game that allowed us to collect large-scale behavioral telemetry and psychological survey instruments in-game from thousands of players, facilitating a better understanding of psychological functioning within the game environment itself (Vuorre et al., 2024; Vuorre et al., 2023).

Why care about digital footprints?

The above examples have highlighted the scientific uses to which digital footprint data can be put: Without unbiased and accurate data on online behaviors, hopes for estimating unbiased causal effects, and for providing an accurate understanding of lives online, are unlikely to be fulfilled. More generally, "digital wellbeing" tools released by platforms like Meta (Meta, 2022), Google (Google, 2022), Apple (Apple, 2018), and TikTok (TikTok, 2021) highlight not only the industry's responses to widespread concerns over these platforms' influence, but also interest among the general public in better understanding their own digital behaviors.

Moreover, current internet technologies and virtual worlds provide their users with affordances that can be too subtle or complicated to self-report on, such as long-term temporal patterns of engagement. An exclusive focus on self-reports could not provide any insight on such behavioral patterns or their consequences. Objective digital footprint data, on the other hand, can shed light on topics, experiences, and behaviors that are beyond individuals' abilities to self-report on, or researchers' abilities to measure otherwise. An increased focus on digital footprints data might yield important or surprising findings on topics and questions that are likely to be beyond the reach of self-report methodologies.

Finally, digital footprints data, and particularly the data donation methods outlined above, enables users to participate in transparent and credible citizen science. Scientific examination of data that are knowingly contributed by individuals empowered with agency and ownership over their digital footprints could lead to a greater scientific understanding among the general population and perhaps inspire future generations of social scientists.

Future directions and conclusions

Future directions

Enrich existing cohort studies with participants digital footprints data

Worldwide, several cohort studies—for example, the UK Biobank and the Adolescent Brain Cognitive Development Study in the US—have collected rich longitudinal data on large cohorts' behavioral, health, educational, and other attributes. These data could be valuable for understanding how digital technologies affect individuals were they to be paired with the participants' digital footprints data.

Be aware of limitations of data alone

Data alone, be it rich behavioral telemetry or broad self-reports, is inert without a good question or theory about what caused it (Pearl and Mackenzie, 2018). The promises of large data can appear illusory, and since it is easy to find questions and their answers within the same data, we must remain cautious with future uses of digital footprints data. Therefore, these data should be subjected to interdisciplinary, and above all, open and transparent examination to avoid fooling ourselves and each other.

Support basic science and methodological innovation within the sciences

As discussed above, maximizing the return on investment of studying digital footprints data requires giving scientists the time and resources needed to carefully develop the required theories and methodologies. To this end, funding mechanisms should not shy away from funding basic theory- and curiosity-driven work, or from purely methodological work such as developing online platforms that improve on and connect APIs and other data donation mechanisms.

Conclusion

The statement "data is the new oil", attributed to the mathematician Clive Humby, aptly summarizes the large-scale data collection efforts of large internet platforms in the past two decades for product development and monetizing. More recently, many social scientists interested in the human effects of these platforms and technologies have increasingly recognized the

potential value of those data for examining platform users' behavior for purposes of scientific understanding for the common good. In this vein, many efforts are under way to use these digital footprints data to examine how technologies affect their users. Yet, to support these efforts, we need to better motivate—and examine the legal and ethical ramifications of doing so—industry stakeholders and societal partners to engage in transparent and credible investigations of our digital footprints.

References

Anderson, M., and J. Jiang, *Teens, Social Media and Technology 2018*, Pew Research Center, May 31, 2018.

Appel, M., C. Marker, and T. Gnambs, 'Are Social Media Ruining Our Lives? A Review of Meta-Analytic Evidence', *Review of General Psychology*, Vol. 24, No. 1, March 2020, pp. 60–74.

Apple, 'iOS 12 Introduces New Features to Reduce Interruptions and Manage Screen Time', Apple Newsroom (United Kingdom), 2018. https://www.apple.com/uk/newsroom/2018/06/ios-12-introduces-new-features-to-reduce-interruptions-and-manage-screen-time/

BBC News, 'China Cuts Children's Online Gaming to One Hour', BBC News, August 30, 2021, sec. Technology.

Best, P., R. Manktelow, and B. Taylor, 'Online Communication, Social Media and Adolescent Wellbeing: A Systematic Narrative Review', *Children and Youth Services Review*, Vol. 41, 2014, pp. 27–36.

Dickson, K., M. Richardson, I. Kwan, W. MacDowall, H. Burchett, C. Stansfield, G. Brunton, K. Sutcliffe, and J. Thomas, 'Screen-Based Activities and Children and Young People's Mental Health and Psychosocial Wellbeing: A Systematic Map of Reviews', 2019.

Ellis, D.A., 'Are Smartphones Really That Bad? Improving the Psychological Measurement of Technology-Related Behaviors', *Computers in Human Behavior*, Vol. 97, August 1, 2019, pp. 60–66.

Ellis, D.A., B.I. Davidson, H. Shaw, and K. Geyer, 'Do Smartphone Usage Scales Predict Behavior?', *International Journal of Human-Computer Studies*, Vol. 130, October 1, 2019, pp. 86–92.

Ghai, S., L. Fassi, F. Awadh, and A. Orben, 'Lack of Sample Diversity in Research on Adolescent Depression and Social Media Use: A Scoping Review and Meta-Analysis', *Clinical Psychological Science*, February 7, 2023, p. 21677026221114859.

Ghai, S., L. Magis-Weinberg, M. Stoilova, S. Livingstone, and A. Orben, 'Social Media and Adolescent Well-Being in the Global South', *Current Opinion in Psychology*, Vol. 46, August 1, 2022, p. 101318.

Google, 'Digital Wellbeing through Technology | Google', Google Digital Wellbeing, 2022. https://wellbeing.google/

Hancock, J., S.X. Liu, M. Luo, and H. Mieczkowski, 'Psychological Well-Being and Social Media Use: A Meta-Analysis of Associations between Social Media Use and Depression, Anxiety, Loneliness, Eudaimonic, Hedonic and Social Well-Being', *SSRN Scholarly Paper, Social Science Research Network*, Rochester, NY, March 9, 2022.

International Telecommunications Union, Facts and Figures 2024, 2024.

Johannes, N., M. Vuorre, and A.K. Przybylski, 'Video Game Play Is Positively Correlated with Well-Being', *Royal Society Open Science*, Vol. 8, No. 2, February 17, 2021, p. 202049.

Kaye, L.K., A. Orben, D.A. Ellis, S.C. Hunter, and S. Houghton, 'The Conceptual and Methodological Mayhem of "Screen Time"', *International Journal of Environmental Research and Public Health*, Vol. 17, No. 10, January 2020, p. 3661.

Meta, 'New Tools and Resources for Parents and Teens in VR and on Instagram', Meta, June 14, 2022. https://about.fb.com/news/2022/06/tools-for-parents-teens-vr-and-instagram/

Meta Platforms, 'User Guide for the Instagram Data Access Pilot for Well-Being Research', Meta Platforms, August 2024.

National Academies of Sciences, *Social Media and Adolescent Health*, Edited by S. Galea, G.J. Buckley, and A. Wojtowicz, National Academies Press, Washington, D.C., 2024.

Odgers, C.L., 'The Great Rewiring: Is Social Media Really behind an Epidemic of Teenage Mental Illness?', *Nature,* Vol. 628, No. 8006, March 29, 2024, pp. 29–30.

Ophir, Y., Y. Lipshits-Braziler, and H. Rosenberg, 'New-Media Screen Time Is Not (Necessarily) Linked to Depression: Comments on Twenge, Joiner, Rogers, and Martin (2018)', *Clinical Psychological Science*, Vol. 8, No. 2, March 2020, pp. 374–378.

Orben, A., 'Teenagers, Screens and Social Media: A Narrative Review of Reviews and Key Studies', *Social Psychiatry and Psychiatric Epidemiology*, Vol. 55, No. 4, April 1, 2020, pp. 407–414.

Parry, D.A., B.I. Davidson, C.J.R. Sewall, J.T. Fisher, H. Mieczkowski, and D.S. Quintana, 'A Systematic Review and Meta-Analysis of Discrepancies between Logged and Self-Reported Digital Media Use', *Nature Human Behaviour*, May 17, 2021, pp. 1–13.

Pearl, J., and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*, First edition., Basic Books, New York, 2018.

Podsakoff, P.M., N.P. Podsakoff, L.J. Williams, C. Huang, and J. Yang, 'Common Method Bias: It's Bad, It's Complex, It's Widespread, and It's Not Easy to Fix', *Annual Review of Organizational Psychology and Organizational Behavior*, Vol. 11, No. Volume 11, 2024, January 22, 2024, pp. 17–61.

Sewall, C.J.R., T.M. Bear, J. Merranko, and D. Rosen, 'How Psychosocial Well-Being and Usage Amount Predict Inaccuracies in Retrospective Estimates of Digital Technology Use', *Mobile Media & Communication*, Vol. 8, No. 3, September 1, 2020, pp. 379–399.

Sewall, C.J.R., and D.A. Parry, 'The Role of Depression in the Discrepancy Between Estimated and Actual Smartphone Use: A Cubic Response Surface Analysis', *Technology, Mind, and Behavior*, Vol. 2, No. 2, July 15, 2021.

The Chosun Daily, 'New Law to Limit Game Times for Young Teens', The Chosun Daily, February 22, 2024. https://www.chosun.com/english/national-en/2011/04/21/OHF2QGKTJGZQ5YRUMGZJVASCSA/

TikTok, 'Digital Well-Being', TikTok, March 5, 2021. https://www.tiktok.com/safety/en/well-being/

Twenge, J.M., T.E. Joiner, M.L. Rogers, and G.N. Martin, 'Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates Among U.S. Adolescents After 2010 and Links to Increased New Media Screen Time', *Clinical Psychological Science*, November 14, 2017, p. 2167702617723376.

Valkenburg, P.M., I. Beyens, J.L. Pouwels, I.I. van Driel, and L. Keijsers, 'Social Media Use and Adolescents' Self-Esteem: Heading for a Person-Specific Media Effects Paradigm', *Journal of Communication*, Vol. 71, No. 1, February 1, 2021, pp. 56–78.

Valkenburg, P.M., A. Meier, and I. Beyens, 'Social Media Use and Its Impact on Adolescent Mental Health: An Umbrella Review of the Evidence', *Current Opinion in Psychology*, Vol. 44, April 1, 2022, pp. 58–68.

Vuorre, M., N. Ballou, T. Hakman, K. Magnusson, and A.K. Przybylski, 'Affective Uplift During Video Game Play: A Naturalistic Case Study', *ACM Games*, Vol. 2, No. 3, August 30, 2024, p. 23:1-23:14.

Vuorre, M., N. Johannes, K. Magnusson, and A.K. Przybylski, 'Time Spent Playing Video Games Is Unlikely to Impact Well-Being', *Royal Society Open Science*, Vol. 9, No. 7, July 27, 2022, p. 220411.

Vuorre, M., K. Magnusson, N. Johannes, J. Butlin, and A.K. Przybylski, 'An Intensive Longitudinal Dataset of In-Game Player Behaviour and Well-Being in PowerWash Simulator', *Scientific Data*, Vol. 10, No. 1, September 13, 2023, p. 622.

Vuorre, M., D. Zendle, E. Petrovskaya, N. Ballou, and A.K. Przybylski, 'A Large-Scale Study of Changes to the Quantity, Quality, and Distribution of Video Game Play During a Global Health Pandemic', *Technology, Mind, and Behavior*, Vol. 2, No. 4, November 8, 2021.

Wald, R., J.T. Piotrowski, T. Araujo, and J.M.F. van Oosten, 'Virtual Assistants in the Family Home. Understanding Parents' Motivations to Use Virtual Assistants with Their Child(Dren)', *Computers in Human Behavior*, Vol. 139, February 1, 2023, p. 107526.

Wilson, C., 'Emails Reveal How Labor Engineered Event to Support Its Own Teen Social Media Ban', Crikey, November 21, 2024. https://archive.ph/Q795S

4.3. Data Donation as Method: Rethinking Access to Digital Trace Data

Jakob Ohme

Weizenbaum Institute

Abstract

This short report introduces the method of data donation as an emerging approach to access user-centric digital trace data for social science research. Against the backdrop of increasingly restricted platform access, data donation offers a user-driven alternative grounded in the legal right to data under the General Data Protection Regulation (GDPR). The report outlines the conceptual basis of data donation, contrasts it with traditional data access methods, and discusses its growing relevance for studying digital media use, particularly in relation to well-being. Drawing on recent developments and experiences from several research projects, the report highlights both the opportunities and challenges of implementing data donation frameworks. In doing so, it offers practical insights into how user-provided data can open new avenues for digital media research and contribute to the development of a more transparent and participatory data access regime. The report concludes with reflections on future directions and the broader implications for the field.

Highlights

- Data donation leverages users' GDPR rights to provide researchers with platform trace data previously inaccessible.
- The method offers a user-centered alternative to APIs, enabling individual-level insights into media use and effects.
- Case studies show how donated data can reveal what users actually see, like political TikToks during elections.
- A new paradigm in media effects research emerges: combining trace data with self-reports for deeper causal insights.
- Despite technical and ethical challenges, open-source tools and interdisciplinary teams are driving adoption forward.

Introduction

This short report will give an overview of the method of data donation and its implementation in social science research. It will specifically focus on the question of how digital media research, with a specific focus on virtual worlds, can utilize this method. It defines and explains the relevance of data donations, gives the conceptual background, and provides an overview of existing data donation frameworks. The report mentions key challenges and solutions of data donations, explains

how they can be used in digital well-being research, and gives an outlook on future development in this vein of the data accessing regime.

Data donations have become an increasingly relevant method in digital media research. It is called 'donations' as users who utilize their right to data that data processors, such as large online platforms, hold about them, request their own data, and donate it to researchers for scientific purposes. Article 15 of the General Data Protection Regulation (GDPR) of the European Union presents citizens with the possibility that data controllers shall provide a copy of the personal data undergoing processing. While GDPR is mainly known for complicating data collection (such as giving consent to store data to general practitioners), it also presents a new data access regime to users (Ausloos and Veale, 2020). GDPR allows only for users' data access within the European Union. However, Very Large Online Platforms (VLOPs) like Facebook or TikTok offer the option of data takeout from the platforms in countries beyond the EU.

The users 'right to data' has attracted researchers' attention, as it opens a window of opportunity to gather individual user data to an extent that was not possible before. In social and computer science, the notoriously difficult access to platform data has long restricted scrutinized research. Access to platform data, such as content distributed, was often only shared with researchers at the discretion of digital platforms, often with limited outcomes. Moreover, shared data usually remained at the level of public and aggregated data. This led to a plethora of research describing the spread of content (for example, on social media networks like Twitter, now X) but to a lack of research explaining the effects of social media content consumption. For this type of research, individual-level data is necessary, but is difficult to gather with previous modes of data collection, such as APIs or scraping (e.g., Ohme et al., 2024). Users' right to data has opened a pathway to access such individual-level data, and although researchers so far cannot access such data directly, the insights available have led to continued efforts to establish this method.

This report is based on relevant literature published around the method of data donations as well as experiences gathered via three different data donation projects between 2020 – 2025.

Conceptual Background

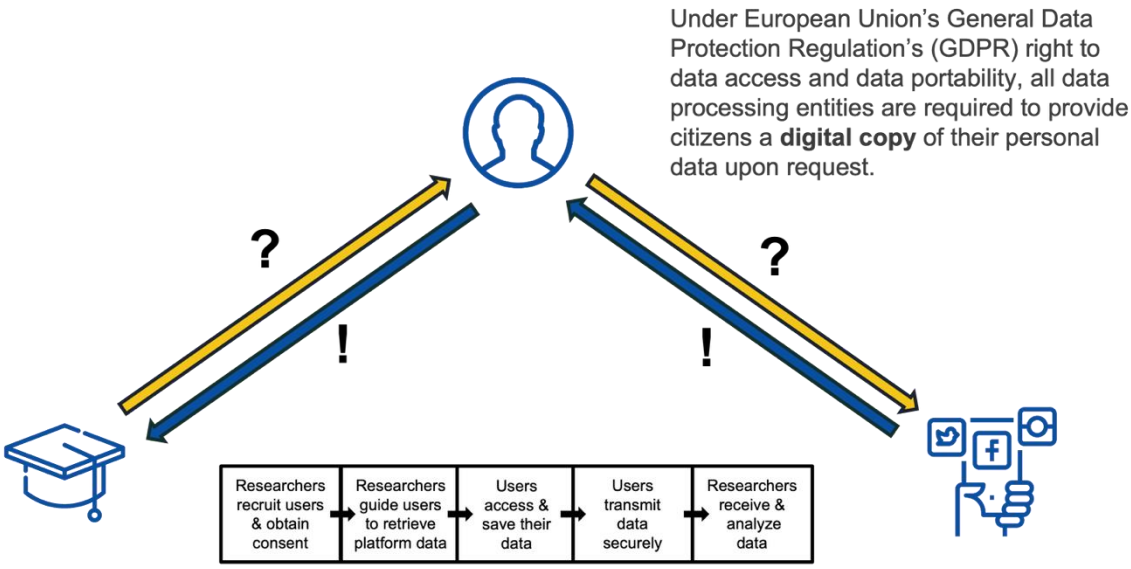
Researcher data access to large online platforms has been historically complicated. Very large online platforms, such as Meta, X, or YouTube, have made access to the data that they process and publish always complicated. At times, researcher data access was easier, for example, with the Twitter Academic API (discontinued in 2023) or Facebook's Crowdtangle (discontinued in 2024). Other initiatives, such as the Social Science One initiative, where Meta worked with selected researchers directly, largely failed to deliver the expected outcomes. With Article 40 of the EU's Digital Services Act (DSA) on researcher data access not being fully implemented yet[8], no reliable methods for data access existed at the time of writing. The discontinuation of free or affordable API solutions came at a time when questions around the role of online platforms for digital well-being,

[8] At the time of writing in June 2025, access to public data (Article 40.12 of the DSA) is partly working, but the European Commission has opened several cases against VLOPs for not fully complying with the regulation. Access to non-public data (Article 40.4 of the DSA), is still waiting for the Delegated Act on this report to be published. Access requests are expected to be filed by early 2026.

democratic discourse, and election integrity became more pressing, necessitating alternative platform data access models.

The GDPR's Transparency Provisions provided a loophole to current restrictions, granting users access to their own data, including data held by data controllers, such as digital platforms, where their data is processed. This led to a fundamental change in the action chain of researcher data access. Rather than researchers requesting access to data from platforms, users could now enable research by requesting and sharing their data, becoming an intermediary between researchers and platforms (see Figure 4).

Figure 4 . Trajectory of the data donation process



Source: Author's own elaboration

User-provided data is important in digital media effects research, mainly for three reasons. First, users can access the data that research needs, thereby playing a central role in the data access process. Second, user-provision changes the type of data, from platform to user-centric. This allows more micro-level media effects research, based on individual digital trace data. Third, with the user in the loop, more transparent consent frameworks are possible, pointing back to the original intention of the GDPR.

Data donations, defined as "donations of existing digital traces with informed consent" (Ohme et al., 2024), utilize all three of these provisions and thereby strongly differ from other data access modes. Compared to access via automated programming interfaces (APIs), which are provided by data controllers and are thereby platform-centric, data donations are user-centric. The data provided is provided retrospectively (e.g., existing digital trace data are collected), relative to tracking approaches (e.g., browser plugins or apps), where digital traces are collected as they are produced, with a prospective timeframe. In data donations, users would rather donate existing data to science, while in tracking approaches, data are produced for science (although potentially as a byproduct of usage). Compared to API or scraping access, users provide explicit consent for their data to be used for research, often even with the possibility to review it before donating. The latter

function is missing in tracking approaches, although white-lists or other specifications might limit the amount of tracked data. API content is often restricted to aggregated-level data that is public and published. Data donations (as well as tracking methods), however, offer non- to semi-public data on an individual level. It is these data that allow for the development of new media effects paradigms (Valkenburg, 2022). The measurement unit of data donations is a user account, which helps device-independent data gathering, while tracking is mostly centered on the device. Compared to other methods, the user involvement for data donations is considerably high. As outlined in section three, several steps are necessary for users to gather and donate their data. User involvement for tracking or APIs is usually lower. Despite the possibility to review the data, the privacy risks of the data donation method are considerable, as the data download packages (DDPs) contain personally identifiable information.

It is essential to state here that narrower and more inclusive understandings of data donations exist. A wider understanding is applied to any type of data in the possession of users that is voluntarily shared for research with scientists. While such a definition could even deem survey responses a 'data donation' (e.g., the opinion of a user becomes data when entering it in a survey tool), the common application restricts the term data donations to digital traces or log data of the user that are gathered by digital services or devices. Here, no restriction to the type of digital service or device exists, as long as user data is gathered and can be requested under data transparency regimes such as, but not restricted to, GDPR. For example, Article 6.9 of the EU's Digital Markets Act on data portability provides a similar provision (e.g., Ausloos & Veale, 2020).

A more narrow approach, sometimes referred to as the 'EU data donation approach', describes the collection of data donations in the format of data download packages (DDPs) through the use of specifically developed software solutions in the form of data donation frameworks. The pipeline used here has developed as a standard, due to the fact that large online platforms such as Meta or TikTok have developed the format of DDPs as a way for users to access their data. In reaction to this, researchers have developed solutions for gathering and processing this type of data. The kind of data provision and the high interest in the large social media companies resulted in this more confined meaning of the data donation method.

Key Challenges and Solutions

Regardless of whether a narrow or inclusive approach is used, the process of collecting data donations remains rather similar.

- Researchers recruit participants and obtain informed consent.
- Participants receive clear instructions on how to access and donate their personal data.
- Participants download their data (e.g., as a Data Download Package or other file format) and store it locally.
- Participants transmit the data to the researchers, ideally after reviewing it and via a secure, privacy-preserving channel.
- Researchers receive the data and proceed with analysis in accordance with ethical and legal standards.

What sounds like a straightforward process is more presuppositional in practice. This section outlines the main challenges and solutions to data donation research.

Data Quality

Challenge: The researcher is dependent on the user (and potentially the platform), so the predictability of content gathered through donations is medium, at best. The structure of the data download package can be determined beforehand, but it might change during the data collection process, resulting in missing data. However, difficulties on the user side to provide the correct content also make this a challenge.

Solution: For data donations with DDPs, platforms need to stop making unannounced changes to their DDPs (e.g., Hase et al., 2024). Researchers need to instruct participants to the best of their ability to help them successfully download their data (e.g., Van Driel, 2022). Software, specifically developed for the purpose, can help to minimize the risk of data quality issues. PORT[9] (developed in the Netherlands) and the Data Donation Module[10] (developed in Switzerland) are two open-source solutions that can help to streamline the process.

Representativeness

Challenge: Every data collection effort that draws a sample is subject to sample biases. Given the high level of user engagement necessary, the chances of collecting data donations from a specific user sample are high. Research so far has shown that minor sample biases towards younger, male, tech-savvy, and privacy-reluctant participants exist (e.g., Wedel et al., 2025; Ohme et al., 2021; Hase & Haim, 2024).

Solution: Minimizing biases through diverse sampling strategies can help; however, there seems to be no way to rule out deviating samples. It is important, however, to compare sample biases of all methods employed to understand the scope of data donation biases. Oversampling specific populations of interest can help secure a sufficient data basis.

Technicality

Challenge: The use of digital trace data in general, and data donations in specific, requires a certain skillset in data science. DPPs do come in machine-readable formats (such as JSON), scraping metadata can be challenging, and content classification is an emerging field where no ready-made solutions exist yet. Hence, working with this method can pose computational challenges to researchers.

Solution: Looking at the examples of successful data donation projects, larger research teams seem to be able to handle these challenges best. Training of researchers in these methods and the subsequent analysis pipeline is also a necessity. Lastly, working in larger data consortiums can help to bundle resources and skills. However, this stands against data minimization initiatives and poses challenges to data sharing across research groups and countries.

Ethics

Challenge: Social science research is especially experiencing a turn towards working with high-sensitivity data. Data donation is a new method; no standards have been developed here. However,

---

9 https://datadonation.eu

10 https://datadonation.uzh.ch/en/

general rules on how to treat personally identifiable data apply. A specific challenge is the processing of the data before consent to donate is given.

Solution: Processing on the local machine of the user is one possible solution (Boeschoten et al., 2022). It is important to work with an institutional review board on the best possible way for participants to give informed consent. Moreover, it is advisable to use a dedicated data donation framework that gives users the opportunity to review the data before they donate it (e.g., Boeschoeten et al., 2022; Pfiffner et al., 2024). Only through these frameworks can users deselect specific data points, minimizing their privacy risk (e.g., Wedel and Ohme, 2025).

How to leverage data donations for media effects research

The donated data can contain a great variety of information. For social media data, often content that users engaged with in the form of liking, sharing, or commenting is part of the DDPs (e.g., Van Driel et al., 2022). For some platforms, the exposure history of watched content in the form of URLs is provided, for example, for TikTok or YouTube (e.g., Wedel et al., 2024). Other data takeouts contain the search histories of users, e.g., for Google. It is beyond the scope of the report to list all the possibilities of existing data. Rather, the following section will describe a few examples of how data donations have been leveraged in research.

In a large-scale data collection effort, researchers from the University of Amsterdam have conducted several studies with adolescents across several months with different waves of measurement. They have used data donations in several ways. In one study, Van Driel and colleagues (2022) show how the use of social media activities across a week differs between adolescents, also describing the coverage issue, with some types of activities occurring rarely, making them less useful for data donation studies (e.g., posting). In another study, they collected the direct messages (DMs) of adolescents from the Instagram DDPs and analysed them with the use of a BERT topic model (Verbej et al., 2024). The findings, for example, show that expressions of happiness and sadness are relatively minor, compared to other topics in the DMs, that adolescents express happiness more often, and that these emotional expressions differ over time. They also find no significant relationship between the emotional expressions and the self-reported well-being of adolescents.

A second, just recently conducted study, in which the author was part of, collected data donations from German TikTok users around the 2025 German Federal Election campaign[11], with the goal to understand the role of short videos in voting decisions. Users were recruited via the project's public and private media partners. While analysis is still ongoing, preliminary results can illustrate the value of data donations. The DDPs of donors contain the URLs of several million videos. Via a TikTok scraper (Bukoldt, 2025), metadata of the videos (such as title, upload date, etc.) can be gathered. In a first step, we then know, for example, if the video was coming from a political party account or was posted by an influencer. In addition, via content classification using Natural Language Processing or multimodal content classification (e.g., Wedel, 2024), the political nature of the videos, their topic, and sentiment can be determined. The DDPs also contain the watch time of every video in a user's individual watch history. Hence, we can know to a rather exact extent, to what content users have been exposed to on TikTok during the election campaign, when, and how

[11] https://dein-feed-deine-wahl.de

long they watched it. These hyper-longitudinal trace data can now be analyzed with different sequence analysis methods (see Fan et al., 2025). By connecting them with self-reported measures of vote decision, such a setup can relate individual user traces to party preference during an election campaign. Preliminary results show, for example, that users mostly see content from the party they voted for, debunking the idea of widespread far-right content being visible for all TikTok users during an election.

This section shows two things: First, data donation enables a unique type of research data that is needed to explain within-participant media effects (e.g., Valkenburg, 2022). Second, to arrive at such results, several additional steps are necessary. In most contexts, data donations need to be augmented, especially if the content of usage is of interest (see Wedel et al., 2024). In other cases, where only log data is of relevance to the research question, more simplified analytical approaches are possible, however (see, for example, Ohme et al., 2021 for smartphone log-data analysis collected via donated screenshots).

Future Directions and Recommendations

Data donation is one of the most promising ways of data collection to arrive at a new digital platform effects paradigm. Despite the existing challenges (see above) and alternative ways of data collection, the method is on its way to establishing itself as a gateway for user-centered media effects research, as it allows for the connection of digital trace data and self-reported outcome variables. However, more efforts are necessary on that path.

So far, only a number of specialized teams are working with this method, not only because of the technical nature of the data collection process. The two existing data donation frameworks are accessible as open-source software, yet institutionalized hosting of these frameworks for other researchers, ideally without additional costs, can help to make this method more accessible. Potentially, an international data donation consortium can facilitate such an endeavor, including training for researchers.

Working in interdisciplinary teams of social and computer scientists and legal researchers has been shown to be promising when facilitating large-scale data donation projects. While social scientists have experience with the study of media effects and the conceptual nature of such research questions, computer scientists can help to successfully gather data and answer such questions, for example, through the automated analysis of data download packages and augmented content structures. The legal framework of the user's right to data has only been applied to a number of digital platforms, mostly from the social media sector. However, the GDPR regulation grants this right to users also for any other data controller within the EU, be it computer game companies or online shopping sites. These are still uncharted territories for data donation research, and legal support can help to clarify which data can be requested in what scope and with what timeline. Working in interdisciplinary teams, hence, seems to be a necessity here.

Several other data access regulations beyond the GDPR’s right to data exist. For example, the Digital Services Act (see Article 40) and the Digital Markets Act (see Article 9) also allow for platform data access for researchers, partly in more direct ways. However, here again, compliance of platforms is necessary, and in recent years, it has become obvious that this compliance is not always guaranteed. In addition, thresholds for researchers to access user data under these access regimes are likely high. The integration of data donation frameworks with other data access frameworks may, hence, be a good way to secure data access, by deciding, for specific study

purposes, which legal framework works best. Moreover, data donation will remain essential to cross-validate platform-centric access modes, such as API requests or data access requests.

Resources are necessary to develop and sustain new data access methods. Given the high value of user-centric data in better explaining outcomes of digital platform use, more dedicated funding schemes are necessary to facilitate a large-scale uptake of the data donation method in media effects research. It is important to acknowledge that such resource-intensive methods are necessary to move the field forward and to establish a new media effects paradigm. Academic institutions, as well as independent funding bodies, should therefore prioritize such methods in their funding decisions.

Conclusion

In retrospect, this short report has illustrated how data donation has emerged as a promising method for advancing digital media research. Grounded in users' "right to data," it enables access to individual-level digital trace data that was previously inaccessible, particularly from large platforms. By shifting the data access chain—placing users as intermediaries—data donations offer a user-centric alternative to traditional platform-dependent methods like APIs or scraping. While challenges around data quality, representativeness, and privacy remain, concrete solutions and frameworks have been developed and successfully implemented. Case studies demonstrate the method's potential for within-subject designs and fine-grained analyses of media exposure and effects. Looking ahead, further institutionalization, interdisciplinary collaboration, and targeted funding are needed to scale this method and integrate it with broader data access regimes. If supported adequately, data donation can serve as a cornerstone of a new media effects paradigm—one that reflects real-world usage patterns and empowers users in the research process.

References

Ausloos, J., and Veale, M., 'Researching with data rights', *Technology and Regulation*, Vol. 2020, Issue 3, 2020, pp. 136–157, https://doi.org/10.26116/techreg.2020.010

Boeschoten, L., Ausloos, J., Möller, J. E., Araujo, T., and Oberski, D. L., 'A framework for privacy preserving digital trace data collection through data donation', *Computational Communication Research*, Vol. 4, Issue 2, 2022, pp. 388–423, https://doi.org/10.5117/CCR2022.2.002.BOES

Bukold, Q., 'TikTok-Content-Scraper' (Version 1.0) [Computer software], 2025, https://www.weizenbaum-library.de/handle/id/814

Fan, Y., Ohme, J., and Wedel, L., 'Exploring temporal dynamics in digital trace data: Mining user-sequences for communication research' (Version 1), 2025, arXiv. https://doi.org/10.48550/ARXIV.2505.18790

Hase, V., Ausloos, J., Boeschoten, L., Pfiffner, N., Janssen, H., Araujo, T., Carrière, T., De Vreese, C., Haßler, J., Loecherbach, F., Kmetty, Z., Möller, J., Ohme, J., Schmidbauer, E., Struminskaya, B., Trilling, D., Welbers, K., and Haim, M., 'Fulfilling data access obligations: How could (and should) platforms facilitate data donation studies?', *Internet Policy Review*, Vol. 13, Issue 3, 2024, https://doi.org/10.14763/2024.3.1793

Hase, V., and Haim, M., 'Can we get rid of bias? Mitigating systematic error in data donation studies through survey design strategies', *Computational Communication Research*, Vol. 6, Issue 2, 1, 2024, https://doi.org/10.5117/CCR2024.2.2.HASE

Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B. B., and Robinson, T. N., 'Digital trace data collection for social media effects research: APIs, data donation, and (screen) tracking. *Communication Methods and Measures*, Vol. 18, Issue 2, 2023, pp. 124–141, https://doi.org/10.1080/19312458.2023.2181319

Ohme, J., Araujo, T., de Vreese, C. H., and Piotrowski, J. T., 'Mobile data donations: Assessing self-report accuracy and sample biases with the iOS Screen Time function', *Mobile Media & Communication*, Vol. 9, Issue 2, 2020, pp. 293–313, https://doi.org/10.1177/2050157920959106

Pfiffner, N., Witlox, P., and Friemel, T. N., 'Data Donation Module: A web application for collecting and enriching data donations', *Computational Communication Research*, Vol. 6, Issue 2, 2024, https://doi.org/10.5117/CCR2024.2.4.PFIF

Valkenburg, P. M., 'Theoretical foundations of social media uses and effects'. In J. Nesi, E. H. Telzer, and M. J. Prinstein (Eds.), *Handbook of Adolescent Digital Media Use and Mental Health* (1st ed.), 2022, pp. 39–60. Cambridge University Press. https://doi.org/10.1017/9781108976237.004

van Driel, I. I., Giachanou, A., Pouwels, J. L., Boeschoten, L., Beyens, I., and Valkenburg, P. M., 'Promises and pitfalls of social media data donations', *Communication Methods and Measures*, Vol. 16, Issue 4, 2022, pp. 266–282, https://doi.org/10.1080/19312458.2022.2109608

Verbeij, T., Beyens, I., Trilling, D., and Valkenburg, P. M., 'Happiness and sadness in adolescents' Instagram direct messaging: A neural topic modeling approach', *Social Media + Society*, Vol. 10, Issue 1, 20563051241229655, 2024, https://doi.org/10.1177/20563051241229655

Wedel, L., Ohme, J., Mayer, A.-T., Fan, Y., and Gaisbauer, F., 'A comparative analysis of data donation behavior across social media platforms', *International Communication Association (ICA) Conference 2025*, 2025, Denver, Colorado, USA.

Wedel, L., and Ohme, J., 'Longitudinal data donation behavior and data omission across four social media platforms', Social Science Open Access Repository (SSOAR), 2025, https://nbn-resolving.org/urn:nbn:de:0168-ssoar-102706-3

Wedel, L., Ohme, J., and Araujo, T., 'Augmenting data download packages – Integrating data donations, video metadata, and the multimodal nature of audio-visual content. *Methods, data, analyses,* 2024, https://doi.org/10.12758/MDA.2024.08

5. Addressing Challenges

5.1. From the AI Act to an XR Act? Assessing EU Policy for XR Safety and Privacy

Emmie Hine

Yale Digital Ethics Centre / University of Bologna / KU Leuven

Abstract

As extended reality (XR) technologies become more prevalent, they bring both opportunities and profound challenges for the protection of fundamental rights in the European Union. XR environments are uniquely immersive, blurring the boundaries between physical and digital spaces and intensifying concerns around safety, privacy, and autonomy. This short contribution examines how six existing EU legislative instruments—the General Product Safety Regulation, the ePrivacy Directive, the General Data Protection Regulation (GDPR), the Digital Services Act (DSA), the Digital Markets Act (DMA), and the Artificial Intelligence (AI) Act—apply to current XR technologies and potential new technological developments. While each framework offers partial safeguards, they might not cover in detail some of the specific affordances and risks of XR. As a result, regulatory gaps persist, particularly in relation to mental and social safety, biometric data use, automated content moderation, and the limits of user consent. The short contribution highlights where interpretation and enforcement of existing laws could be strengthened and outlines targeted areas for reform or new regulation. It also argues that the EU must expand its conceptions of privacy and safety to reflect the distinctive characteristics of immersive environments. Finally, the contribution emphasizes the need for coordinated research, education, and soft law approaches to address longer-term and intersectional risks. By acting early and deliberately, the EU can position itself as a global leader in XR governance and uphold a rights-based digital future.

Highlights

- Assesses how EU laws apply to XR regarding user privacy and safety.
- Identifies key regulatory gaps in immersive XR environments.
- Calls for expanded concepts of privacy and safety in XR regulation.
- Recommends XR-specific updates to existing digital regulation.
- Advocates EU leadership in global XR governance to uphold fundamental rights.

The European Union (EU)'s governing bodies are increasingly concerned with the effects of extended reality (XR) technology and experiences. While XR offers access to novel experiences and the potential to increase access to education and healthcare, it also threatens certain fundamental rights, among them safety and privacy. Because XR technologies are more immersive than traditional computing platforms, the potential impact on people and their fundamental rights is much greater (Hine et al., 2024), and because their data collection is often continuous and involuntary, they challenge the assumptions underlying existing digital regulation—including the

meaning of privacy, consent, and harm. The European Commission is aware of the need to address these risks. However, before developing new policy or legislation, it is important to understand how existing legislation already applies to current and likely future technological development. This assessment is especially timely given growing EU interest in XR innovation and governance, particularly in virtual worlds. This short report summarizes how six areas of legislation—relevant product safety legislation, the ePrivacy Directive, the General Data Protection Regulation (GDPR), the Digital Markets Act (DMA), the Digital Services Act (DSA), and the Artificial Intelligence (AI) Act—apply to XR and what gaps remain to be filled. It draws on several publications, including Hine et al. (2024).

Product safety legislation

Product safety legislation applies to XR equipment under the General Product Safety Directive[12], which was revised in 2021 to address online marketplaces and new technologies. The revision[13], effective December 2024, specifically acknowledges new health risks from technologies, including psychological, developmental, and mental risks, potentially covering physical, mental, and social safety impacts of XR technology. The NIS 2 Directive[14] will require Member States to implement cybersecurity training and mandate that online platforms and marketplaces implement security measures to prevent data breaches, which will improve the security of XR platforms.

The new Product Liability Directive[15], fully effective in 2026, would extend liability to software and digital services, including medically recognized psychological harm. While it would eliminate the limiting €500 property damage threshold, it would not cover non-medical mental health impacts or social harms. The now-withdrawn AI Liability Directive[16] would have protected victims harmed by AI systems in XR platforms by facilitating evidence access and reducing the burden of proving direct causation. Even if this had been enacted, challenges remain in ensuring users can identify when automated systems have caused them harm; further legislation should address this.

---

12 Directive 2001/95/EC of the European Parliament and of the Council of 3 December 2001 on General Product Safety (Text with EEA Relevance) OJ L 11, 15.1.2002, p. 4-17

13 Regulation (EU) 2023/988 of the European Parliament and of the Council of 10 May 2023 on General Product Safety, Amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council and Directive (EU) 2020/1828 of the European Parliament and the Council, and Repealing Directive 2001/95/EC of the European Parliament and of the Council and Council Directive 87/357/EEC (Text with EEA Relevance) OJ L 135, 23.5.2023, p. 1-51

14 Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on Measures for a High Common Level of Cybersecurity across the Union, Amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and Repealing Directive (EU) 2016/1148 (NIS 2 Directive) OJ L 333, 27.12.2022, p. 80-150.

15 Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on Liability for Defective Products, 28.9.2022

16 Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive), 28.9.2022.

ePrivacy Directive

XR equipment likely qualifies as "terminal equipment" under the ePrivacy Directive[17] due to its Internet connectivity, subjecting it to several requirements. Article 5 requires service providers to maintain security and confidentiality while obtaining explicit consent for data storage/access, except when strictly necessary for service provision. However, the directive's protection is limited to data stored on devices, not after transmission.

Under Article 15, Member States can override confidentiality for security purposes, but the CJEU has ruled[18] that broad metadata retention is only proportionate for genuine national security threats. For serious crimes, data retention must have specific links to public security threats. Without clear definitions of valid security threats, controlling surveillance expansion through XR data retention remains challenging.

The proposed ePrivacy Regulation[19] would have extended privacy rules to electronic communications in XR environments and protect machine-to-machine communications, including XR data outside interpersonal communications. While this would have enhanced protection against data interception, the regulation was withdrawn.

GDPR

It is unclear how effectively the GDPR[20] will apply to XR. The GDPR deals with "personal data," defined as "any information relating to an identified or identifiable natural person" (Article 4(1)). The European Parliament briefing on the metaverse acknowledges that the distinction between a data controller and data processor (Articles 24-28) will become blurred, which raises questions about where to collect user consent (Articles 6-7) and display privacy notices (Articles 12-13), especially if data collection will be "involuntary and continuous" (Madiega et al., 2022).

The global nature of VR platforms raises jurisdiction questions, though EU adequacy decisions partially address data transfer issues. The GDPR is noted for the Brussels Effect, and this could transfer to XR platforms, which could default to the strongest protections globally. The GDPR's Article 6 provides various legal bases for data processing, including consent, contract performance, and "legitimate interests." The European Data Protection Board (EDPB) has ruled against using the contract clause for targeted advertising[21], and TikTok's attempt to use "legitimate interests" was

---

17 Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 Concerning the Processing of Personal Data and the Protection of Privacy in the Electronic Communications Sector (Directive on Privacy and Electronic Communications) 2002 OJ L 201, 31.7.2002, p. 37.

18 CJEU, La Quadrature du Net and Others v Premier ministre and Others, judgment of 6 October 2020, joined cases C-511/18, C-512/18 and C-520/18, §136.

19 Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL Concerning the Respect for Private Life and the Protection of Personal Data in Electronic Communications and Repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications), 10.1.2017.

20 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) OJ L 119, 4.5.2016, p. 1-88.

21 "Binding Decision 3/2022 on the Dispute Submitted by the Irish SA on Meta Platforms Ireland Limited and Its Facebook Service (Art. 65 GDPR)" 2022.

challenged (Lomas, 2023). Although seemingly a legitimate justification for data processing, consent dialogues often use deceptive presentation of information and fatigue users with their quantity (Utz et al., 2019). Thus, when presented with an ostensibly valid consent choice, users could end up involuntarily consenting to more or different data collection than they intended to.

Article 9 prohibits processing biometric and sensitive data without explicit consent, with some exceptions including data "manifestly made public." While platforms might argue that using XR in public spaces makes some biometric data public, this interpretation becomes problematic for detailed movement tracking and internal biometric measurements. Article 20's data portability rights could enable XR platform interoperability, though this would require new data standards. Article 22 restricts solely automated decision-making, potentially affecting automated content moderation in XR platforms.

Recent CJEU rulings, including OT v Vyriausioji tarnybinės etikos komisija,[22] may protect against the inference of sensitive information. As aggregated and non-personal data falls outside the GDPR's purview, this ruling could protect XR users from having sensitive inferences made about them without their knowledge, although they remain vulnerable to the use of anonymized or synthetic data based on data to mine behavioral insights at a group level (Renieris, 2023, p.120). Notable cases include fines for excessive employee video monitoring (LfD Lower Saxony, 2021) and rulings against invasive proctoring software,[23] which could set precedents for limiting surveillance in XR environments.

Digital Services Act

As part of the EU's flagship platform regulation package, the DSA[24] regulates illegal content and targeted advertising on digital platforms. It implements a "notice and action" system for content removal, with priority given to "trusted flaggers." Terrorist content must be removed within one hour under the Terrorism Regulation, but content jurisdiction in pan-jurisdictional XR environments remains complex (Hine, 2023). XR platforms contain not just static content but dynamic conduct, which makes automated and human content moderation difficult. Conduct is ephemeral and thus has to be moderated in real time, but at present, most platforms rely primarily on human moderation (Schulenberg et al., 2023; Gray, Carter, and Egliston, 2024), which does not scale. However, immersive content and conduct is difficult to characterize for automated moderation, which also carries the risk of increasing bias.

The DSA prohibits ads targeting minors and using sensitive characteristics, requiring real-time disclosure of advertisers and targeting methods. For "very large online platforms" (VLOPs), it mandates maintaining accessible ad repositories and prohibits manipulative design (or "dark patterns"). While the EDPB has guidelines for identifying dark patterns on social media,[25] these need adaptation for immersive environments.

---

22 Judgement of 1 August 2022, OT v Vyriausioji tarnybinės etikos komisija (2022)

23 "Deliberação/2021/622" 2021.

24 Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act) OJ L 277, 27.10.2022, p. 1-102.

25 "Guidelines 3/2022 on Dark Patterns in Social Media Platform Interfaces: How to Recognise and Avoid Them" 2022.

A significant limitation is that many key requirements, including systemic risk analysis and independent auditing, only apply to platforms with over 45 million monthly EU users. This creates a potential regulatory gap for smaller XR platforms that don't meet the VLOP threshold but could still pose significant risks (Laux, Wachter, and Mittelstadt, 2021).

Digital Markets Act

The DMA[26] regulates large online "gatekeepers" providing core platform services. Virtual worlds may be categorized as "online social networking services" and "online intermediation service providers" (Lopez-Tarruella and Rodríguez de las Heras Ballell, 2024). As with the DSA, the DMA's current applicability is limited because no XR platforms are designated as "core platform services," but some provisions will apply to Apple, ByteDance, and Meta, which are designated gatekeepers and active in XR (European Commission, 2024). For those platforms, Article 5 can help protect user privacy by preventing gatekeepers from combining personal data from their core platform services with non-core platform services or cross-using data across core platform and other services. Non-gatekeepers are not subject to these restrictions. In the future, XR operating systems may be required to allow third-party app stores and platforms will have to facilitate data transfers between platforms and permit hardware and software interoperability (Lopez-Tarruella and Rodríguez de las Heras Ballell, 2024). At present, though, the DMA's applicability to XR platforms and companies is highly limited.

AI Act

The AI Act[27] regulates AI systems in physical and some virtual environments. Relevant requirements for virtual environments include the disclosure of AI interactions, including artificial avatars and synthetic content. Furthermore, users must be notified about emotion recognition and biometric categorization systems.

While real-time biometric identification is banned for law enforcement in physical spaces (with broad exceptions), it remains permitted in virtual environments. Furthermore, law enforcement acting under an exception—or non-law-enforcement-actors—could use AR devices equipped with facial recognition to recognize protestors or other individual in public (Hine, Cowls, and Floridi, 2024). Emotion recognition systems are banned in educational and workplace settings, which would seem to preclude the use of those systems in XR educational and work platforms. Furthermore, profiling based on biometric data to infer protected characteristics is prohibited, though "lawfully applied biometric datasets" in law enforcement are exempt.

The AI Act bans systems that cause "significant harm" through "subliminal techniques beyond a person's consciousness" or "purposefully manipulative or deceptive techniques," potentially

---

[26] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on Contestable and Fair Markets in the Digital Sector and Amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) OJ L 265, 12.10.2022, p. 1-66.

[27] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) OJ L, 2024/1689, 12.7.2024

protecting users in XR environments. However, there is little guidance for what constitutes "significant harm" or "subliminal techniques", especially when it comes to virtual environments. While emotion recognition in work and educational contexts is banned, AI-based performance monitoring systems are allowed as "high-risk," which could permit employer monitoring of XR work devices. Similarly, facial image scraping is prohibited, but collecting non-facial biometric data through XR devices remains permitted under high-risk classification.

Recommendations

Detailed recommendations can be found in Hine et al. (2024). This short report will highlight several broad categories of recommendations. First, on a fundamental level, privacy should not be construed as only data protection. XR technologies implicate decisional and local privacy as well—the right to make decisions without interference and the right to have a space where one can "just be" without observation (Hine et al., 2024; Roessler, 2018). Safety should also be expanded beyond just physical safety; mental and social safety are crucial to the holistic wellbeing of citizens (Hine et al., 2024). New product safety legislation in the EU is increasingly recognizing this.

In terms of concrete recommendations, additional safety measures should be mandated to protect XR users from harassment, and assault and battery laws should be clarified to cover virtual attacks where no physical contact occurs. Age verification should be mandated at the account and device level. DSA and DMA provisions on advertising, dark patterns, data processing, and portability should be expanded to all XR platforms, and advertisement archives should have additional detail to cover where and how an ad was displayed or performed in an XR environment. The AI Act should be interpreted to ban targeted transitive and subliminal advertising to protect user autonomy, and the DSA to require effective automated and human content moderation. The scraping of any form of biometric data and the nonconsensual aggregation of biometric data should be banned by legislation. Additionally, the GDPR should be analyzed to determine if the data processor/controller distinction is still fit for purpose (Martin, 2022).

Not all recommendations require creating or modifying hard law; much can be accomplished with soft law and other actions. EU Member States should also fund research into the long-term impacts of XR, promote research and initiatives around safe drinking habits in XR (anecdotal reports suggest alcohol misuse may be a significant problem in VR), support XR literacy campaigns, and study the impacts of harassment and counteracting measures. This research should specially examine impacts on marginalized and disabled users, as well as children.

Conclusion

XR will challenge the EU digital legislation landscape, but new technologies have emerged before, and the law has adapted to account for them. Currently, the risks of XR to safety and privacy can be managed primarily with careful interpretation and potential amending of existing laws. Yet XR technologies push at the conceptual limits of these laws—demanding a shift from narrow notions of data protection to more expansive ideas of privacy, autonomy, and mental and social safety. As we bring more rights into consideration, such as freedoms of assembly and expression (Hine, Cowls, and Floridi, 2025), we should not foreclose the idea of new legislation along the lines of the AI Act. Furthermore, global harmonization of XR governance will be crucial because XR involves companies and users from across the globe interacting in shared digital spaces. By acting early, the EU can not only anticipate future legal fragmentation, but also shape the global conversation around

immersive rights governance—positioning itself as a leader in building XR systems that are safe, equitable, and respectful of fundamental rights.

References

'Binding Decision 3/2022 on the Dispute Submitted by the Irish SA on Meta Platforms Ireland Limited and Its Facebook Service (Art. 65 GDPR)', December 5, 2022. https://edpb.europa.eu/our-work-tools/our-documents/binding-decision-board-art-65/binding-decision-32022-dispute-submitted_en

'Deliberação/2021/622', Comissão Nacional de Proteção de Dados, May 28, 2021.

Directive 2001/95/EC of the European Parliament and of the Council of 3 December 2001 on General Product Safety, OJ L, Vol. 011, Vol. 011, 3.12.2001.

Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 Concerning the Processing of Personal Data and the Protection of Privacy in the Electronic Communications Sector (Directive on Privacy and Electronic Communications), OJ L 201, Vol. 201, Vol. 201, 31.7.2002.

Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on Measures for a High Common Level of Cybersecurity across the Union, Amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and Repealing Directive (EU) 2016/1148 (NIS 2 Directive), OJ L 333, 27.12.2022.

European Commission, 'DMA Designated Gatekeepers', May 13, 2024. https://digital-markets-act.ec.europa.eu/gatekeepers_en

Gray, J.E., M. Carter, and B. Egliston, 'Trust and safety in social VR: Current industry practices', in J.E. Gray, M. Carter, and B. Egliston (eds.), *Governing social virtual reality: Preparing for the content, conduct and design challenges of immersive social media*, Springer Nature Switzerland, Cham, 2024, pp. 61–75.

'Guidelines 3/2022 on Dark Patterns in Social Media Platform Interfaces: How to Recognise and Avoid Them', European Data Protection Board, March 14, 2022.

Hine, E., 'Content moderation in the Metaverse could be a new frontier to attack freedom of expression', *Philosophy & Technology*, Vol. 36, No. 3, June 16, 2023, p. 43.

Hine, E., J. Cowls, and L. Floridi, 'Assembly and expression in extended reality: Transposing fundamental rights across realities', *Proceedings of the International Congress Towards a Responsible Development of the Metaverse*, Alicante, 2024.

———, 'Assembly and expression in extended reality: Transposing human rights across realities', *Interactive Entertainment Law Review*, Vol. 8, No. 1, 2025, pp. 66–80.

Hine, E., I.N. Rezende, H. Roberts, D. Wong, M. Taddeo, and L. Floridi, 'Safety and privacy in immersive extended reality: An analysis and policy recommendations', *Digital Society*, Vol. 3, No. 2, July 3, 2024, p. 33.

La Quadrature Du Net and Others v Premier Ministre and Others, ECJ 2020.

Laux, J., S. Wachter, and B. Mittelstadt, 'Taming the few: Platform regulation, independent audits, and the risks of capture created by the DMA and DSA', *Computer Law & Security Review*, Vol. 43, November 1, 2021, p. 105613.

LfD Lower Saxony, 'LfD Lower Saxony Imposes a Fine of 10.4 Million Euros on Notebooksbilliger.De', January 8, 2021. https://lfd.niedersachsen.de/startseite/infothek/presseinformationen/lfd-niedersachsen-verhangt-bussgeld-uber-10-4-millionen-euro-gegen-notebooksbilliger-de-196019.html.

Lomas, N., 'Meta tries to keep denying EU users a free choice over tracking -- but change is coming', TechCrunch, March 30, 2023. https://techcrunch.com/2023/03/30/meta-facebook-gdpr-ads-tracking/

Lopez-Tarruella, A., and T. Rodríguez de las Heras Ballell, 'A European regulatory framework for the Metaverse' MetaverseUA Chair Research Paper #1', SSRN Scholarly Paper, Social Science Research Network, Rochester, NY, November 16, 2024.

Madiega, T., P. Car, M. Niestadt, and L. Van de Pol, *Metaverse: Opportunities, Risks and Policy Implications*, June 2022.

Martin, B., 'Privacy in a programmed platform: How the General Data Protection Regulation applies to the Metaverse', *Harvard Journal of Law & Technology*, Vol. 36, Issue 1, 2022.

OT v Vyriausioji Tarnybinės Etikos Komisija, ECJ 2022.

Proposal for a Directive of the European Parliament and of the Council on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive), 28.9.2022.

Proposal for a Directive of the European Parliament and of the Council on Liability for Defective Products, 28.9.2022.

Proposal for a Regulation of the European Parliament and of the Council Concerning the Respect for Private Life and the Protection of Personal Data in Electronic Communications and Repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications), 10.1.2017.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), OJ L, Vol. 119, Vol. 119, 4.5.2016.

Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on Contestable and Fair Markets in the Digital Sector and Amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act), OJ L, Vol. 265, Vol. 265, 12.10.2022.

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act), OJ L, Vol. 277, Vol. 277, 27.10.2022.

Regulation (EU) 2023/988 of the European Parliament and of the Council of 10 May 2023 on General Product Safety, Amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council and Directive (EU) 2020/1828 of the European Parliament and the Council, and Repealing Directive 2001/95/EC of the European Parliament and of the Council and Council Directive 87/357/EEC, OJ L, Vol. 135, Vol. 135, 10.5.2023.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA Relevance), 13.6.2024.

Renieris, E.M., *Beyond Data: Reclaiming Human Rights at the Dawn of the Metaverse*, MIT Press, 2023.

Roessler, B., 'Three Dimensions of Privacy'*, Handbook of Privacy Studies: An Interdisciplinary Introduction*, Amsterdam University Press, Amsterdam, 2018, pp. 137–142.

Schulenberg, K., L. Li, G. Freeman, S. Zamanifard, and N.J. McNeese, 'Towards Leveraging AI-Based Moderation to Address Emergent Harassment in Social Virtual Reality', *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ACM, Hamburg Germany, 2023, pp. 1–17.

Utz, C., M. Degeling, S. Fahl, F. Schaub, and T. Holz, '(Un)Informed Consent: Studying GDPR Consent Notices in the Field', *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, ACM, London United Kingdom, 2019, pp. 973–990.

5.2. Content Moderation in the Metaverse: Legal Framework and Practical Challenges

Julián López Richart

Universidad de Alicante

Abstract

This short report explores the legal and practical challenges of content moderation in the metaverse, a conceptually evolving digital ecosystem characterized by immersive, synchronous, and often ephemeral interactions. Unlike traditional Web 2.0 platforms where moderation mechanisms have matured around static text and audio-visual content, the metaverse demands new approaches due to its real-time, embodied nature. Harmful behaviours—such as harassment or manipulation—can occur through gestures, voice, and proximity, making them harder to detect and address using existing moderation tools. The report further examines the applicability of the EU Digital Services Act (DSA) to virtual environment platforms. The DSA establishes a tiered regime of due diligence obligations applicable to intermediary service providers. A significant portion of these obligations aims to regulate content moderation carried out by online platforms, insofar as such moderation may affect fundamental rights, such as freedom of expression and access to information. The DSA does not dictate which content platforms may allow or remove, except in cases where the content is manifestly illegal and the provider has been made aware of its presence. Rather, it establishes a framework of obligations aimed at ensuring transparency and accountability in platform decision-making. While these provisions, in principle, extend to virtual worlds, their effectiveness is limited by the fact that the DSA was primarily designed with Web 2.0 platforms in mind. Immersive and ephemeral interactions typical of metaverse environments pose significant challenges to tracking misconduct and implementing effective appeal mechanisms. Additionally, enforcement becomes more complex in decentralized platforms, where governance is distributed among user communities rather than vested in a single legal entity.

Highlights

- Content moderation faces new legal and technical challenges in the metaverse.
- Immersive, real-time interactions complicate traditional moderation models.
- The DSA provides a useful but incomplete framework for virtual environments.
- Balancing user rights with safety is key for a responsible development of the metaverse.

Introduction

In Web 2.0, the challenge of moderating user-generated content (UGC) on platforms like Facebook, Twitter, and YouTube already strained the boundaries of existing legal frameworks. Faced with the proliferation of harmful content—ranging from hate speech and misinformation, violations of privacy and intellectual property, terrorist content or child pornography—technology companies

implemented mechanisms to detect and react against illegal content or content that they consider harmful (lawful but awful). These content moderation policies reflect both the economic incentives of platforms to maintain safe user environments (Klonick, 2018) and the regulatory pressure from governments, which, as noted in Recital 59 of Directive 2001/29/CE, view these providers as best positioned to respond swiftly and effectively to illegal and harmful online content.

While traditional Web 2.0 platforms are predominantly built around text, images, and videos—content that can be filtered out, removed, or flagged using different well-established techniques—interactions through virtual reality worlds are spatial, embodied, and temporally fluid. Users may attend a concert, explore a museum, participate in a work meeting, or experience intimate conversations in real time, all within shared digital spaces. Harmful conduct in virtual worlds may involve verbal abuse through voice chat, gestures through avatar movements, or psychological manipulation in persistent environments. Such actions often escape the logic of static 'content' as previously understood, raising the stakes for effective and rights-respecting moderation.

This short report examines the legal and practical challenges of moderating content and behaviour in the metaverse, with particular attention to the European Union's regulatory framework. It critically examines the applicability of the Digital Services Act (DSA), explores emerging moderation models tailored to immersive environments, and addresses the growing tensions between governance structures, decentralization, user rights, and the technological features of these virtual spaces.

The Metaverse as a regulatory challenge

The term metaverse—in its singular form—typically denotes a unified, persistent, and interoperable digital ecosystem that integrates physical and virtual realities into a continuous, immersive environment. Conceptually, it aspires to function as an expansive digital infrastructure wherein users, avatars, assets, and experiences can move seamlessly across platforms and contexts (Ball, 2022). In contrast, metaverses or virtual worlds in the plural refer to the multitude of discrete, often self-contained digital environments that already exist today (European Commission, 2023). These include platforms such as Second Life, Roblox, or VRChat, each with its own technological architecture, user base, and internal logic. Such virtual worlds may offer immersive experiences, yet they lack the interoperability and shared continuity that define the singular metaverse ideal. Thus, the metaverse represents an overarching paradigm—still largely aspirational—that envisions their convergence into a cohesive and interconnected digital universe. However, for the sake of conceptual clarity and simplicity, this report will use the singular term 'metaverse' as a shorthand to collectively refer to these diverse and currently fragmented digital environments.

The metaverse's most disruptive features are those that most profoundly challenge existing legal norms: immersion, synchronicity, and persistence (López-Tarruella and Rodríguez de las Heras, 2024). Traditional platforms allow asynchronous interaction. One user posts a message or video; another responds hours later. In that context, automated mechanisms may be used to filter out certain types of content that meet specific parameters—for example, content containing particular keywords, images matching known hashes of illegal material, or patterns indicative of coordinated disinformation campaigns. Moreover, moderation in the Web 2.0, whether carried out by humans or automated means, can be delayed without fundamentally altering the platform's dynamics. In contrast, the metaverse thrives on synchronous interaction. A user's gesture, voice, or proximity may elicit an immediate reaction, and any failure to prevent such conduct in real time may result in emotional harm or reputational damage.

Far from being a mere extension of social media, virtual reality environments represent a paradigm shift in how we communicate, work, learn, and entertain ourselves. In addition, the psychological impact of immersive environments intensifies users' experiences (Hine, 2023; Freeman et al., 2020; Jurecic Rozenshtein, 2021). Studies show that users identify closely with their avatars, attributing actions and events in virtual spaces to real-world consequences. Instances of virtual sexual assault or racial abuse, though digital in form, can evoke trauma akin to their physical-world counterparts (Wiederhold, 2022), highlighting the seriousness of such behaviour. Indeed, incidents of inappropriate conduct have already triggered public concern and calls for stronger governance.

Another factor to take into consideration is the diversity of actors and platforms involved in the development of the metaverse. Unlike the consolidated Web 2.0 space dominated by a handful of tech giants, the metaverse encompasses a wide range of initiatives, including corporate-led platforms such as Meta's Horizon Worlds, open-world environments like Roblox, and decentralized virtual worlds such as Decentraland and The Sandbox. These differing architectures influence how moderation is implemented, who holds decision-making power, and what remedies are available to users (López-Tarruella and Rodríguez de las Heras, 2024).

Existing typologies of content moderation and their limitations in the metaverse

Since the early 2000s, the proliferation of user-generated content across digital platforms has rendered content moderation an increasingly complex and multidimensional task. Moderation mechanisms must balance the protection of fundamental rights, such as freedom of expression, with the imperative to prevent illegal or harmful behaviours. To this end, platforms have developed a variety of moderation models and sanctions against users who violate community guidelines, ranging from content removal and temporary suspensions to permanent bans and algorithmic downranking. However, in the context of immersive virtual environments, implementing effective mechanisms for monitoring user behaviour and address inappropriate conducts presents distinct challenges.

A considerable number of user interactions within these spaces are synchronous and ephemeral, rendering traditional moderation strategies inadequate. Major virtual reality platforms still depend heavily on human moderators—whether company employees, contracted workers, creators of virtual spaces, or even users themselves (Schulenberg et al., 2023). While manual moderation can offer greater accuracy and sensitivity to context, it does not scale well in environments where millions of interactions occur. Effective moderation in these settings often requires real-time responses and the 'physical' presence of moderators within the virtual world to intervene quickly and prevent harmful or inappropriate behaviour. In practice, however, moderators face considerable challenges. Their ability to observe user behaviour is limited, as their sensory perception is designed to simulate real-world constraints. As a result, it is difficult to monitor all activity. Additionally, the ephemeral nature of many interactions often forces moderators to rely on user reports without sufficient tools to verify what actually took place (Sabri et al., 2023). To address these limitations, one potential approach involves enhancing moderators' oversight capabilities and/or implementing comprehensive recording of activities across all areas of the metaverse, thereby enabling the retrospective review of potentially illicit behaviour (López-Tarruella and Rodríguez de las Heras, 2024). However, such measures raise significant concerns regarding user privacy and data protection. Moreover, the pervasive surveillance implied by continuous monitoring may inhibit user spontaneity, as individuals are likely to alter their behaviour if they are aware that their actions are constantly observed or recorded (Castro, 2022).

Automated moderation systems, which rely on code to detect and remove content based on predefined criteria, also face significant limitations in virtual reality environments. Although these systems can be effective for filtering text or identifying known visual content—such as child sexual abuse material—they are not well suited to interpreting the nuances of language, contextualizing human interactions, or discerning the intent behind users' behaviour (Singh, 2019). As a result, they may fail to detect subtle forms of abuse or misinterpret benign interactions, highlighting the need for more sophisticated and context-aware approaches to moderation in immersive settings.

The effectiveness of traditional measures to address illegal conducts or breaches of platform rules is also called into question in the context of immersive environments and virtual worlds. In conventional digital platforms—where user interactions primarily involve text (e.g., posts, comments) or audio-visual content (e.g., images, videos, music) that is stored and remains permanently accessible—standard moderation tools such as content removal or visibility reduction may be effective for managing harmful or unlawful material. However, the metaverse introduces a fundamentally different mode of interaction: users engage with one another through avatars in shared, real-time, multisensory experiences. This shift challenges the adequacy of traditional moderation frameworks. In the metaverse, the focus is not solely on content moderation, but on regulating behaviours that unfold within specific spatial and temporal settings. These behaviours can be highly context-dependent and ephemeral, making them difficult to detect and interrupt in real time (Castro, 2022). As a result, existing moderation mechanisms may need to be rethought or adapted to effectively address the unique characteristics of immersive virtual environments.

One way to address illicit or inappropriate behaviour in the metaverse is by adopting preventive measures that mirror sanctions imposed by judicial or administrative authorities in the physical world. These may include restricting access to specific virtual spaces, limiting user mobility, issuing virtual restraining orders, or disabling the use of certain digital objects. This approach builds on what some scholars describe as 'governance-by-the-metaverse', which refers to the underlying code and rules that structure user behaviour in virtual environments (Janssen, 2022). For instance, although Horizon Worlds initially offered a 'Safe Zone' feature that allowed users to create a protective bubble around their avatars, Meta introduced a new default tool called 'Personal Boundary' in response to early reports of virtual sexual harassment. This feature prevents unwanted physical proximity by making an avatar's hands disappear if they come too close to another avatar, unless explicitly allowed.

Such measures must be carefully balanced to avoid undermining the user experience and to preserve the metaverse as a space for authentic social interaction, similar to the physical world. Moreover, increasing preventive controls based on predictive assessments of potential misconduct raises serious concerns about users' rights. It is therefore essential to weigh the potential harm of such measures aim to prevent damages against the rights and freedoms they may infringe (López-Tarruella and Rodríguez de las Heras, 2024). In cases where restrictions are imposed as a sanction for misconduct, it is also important to uphold legal principles such as proportionality, prior specification of punishable behaviour, and due process.

Another challenge of virtual spaces is adapting content moderation to the legal requirements of different countries—a common practice among current social media platforms. For example, many YouTube videos are restricted in certain regions due to the territorial nature of copyright law, which grants content owners exclusive rights that vary by jurisdiction. As described by the Court of Justice in Case C-507/17 (Google), two approaches—or a combination thereof—are commonly employed to address this issue. First, because platforms often maintain country-specific sites (e.g., Facebook.es, Facebook.fr, Facebook.it), they can selectively remove content in some regions but not others.

Second, geo-blocking techniques may be used to prevent users in a particular territory from accessing certain material. However, applying such geographically based restrictions in the metaverse may conflict with its immersive nature, as it could lead to users physically located in different countries but sharing the same virtual space to experience entirely different realities—hearing different music or seeing different objects—thereby undermining a unified sense of presence and shared experience (Hine, 2023).

Content moderation in the metaverse under the Digital Services Act

The Digital Services Act (DSA), adopted in 2022, is the EU's most ambitious effort to harmonise rules for digital intermediaries. Building upon the e-Commerce Directive of 2000, the DSA introduces a tiered regulatory framework based on the nature and size of the service provider.

At its core, the DSA maintains the principle that hosting service providers are not liable for user content unless they have actual knowledge of its illegality and fail to act promptly. An important clarification is introduced in Article 7, ensuring that voluntary moderation does not negate liability protections ('Good Samaritan clause').

Unlike the e-Commerce Directive, the Digital Services Act (DSA) not only seeks to ensure the smooth functioning of the internal market and foster innovation through a broad liability exemption regime, but also aims to establish harmonised rules that promote a safe, predictable, and trustworthy online environment—one in which the fundamental rights enshrined in the Charter of Fundamental Rights of the European Union are effectively safeguarded (Art. 1.1 DSA). Achieving this objective requires intermediary service providers to act responsibly and with due diligence (Recital 3 DSA). Consequently, a significant portion of the DSA is dedicated to regulating the due diligence obligations imposed on these providers, some of which pertain directly to content moderation practices.

In essence, the DSA recognises the freedom of intermediary service providers to develop and implement their own content moderation policies (Rodríguez de las Heras, 2023). However, it also establishes a set of obligations designed to ensure transparency, proportionality, and accountability in the decisions they make. Within this framework, the role of courts and supervisory authorities is not to determine what content should or should not be removed from a platform, but rather to ensure compliance with minimum standards regarding the safeguards afforded to users, an idea that had been already eloquently defended in the literature (Citron, 2008).

Intermediary service providers are required to clearly outline in their terms and conditions any limitations placed on the use of their services (Art. 14 DSA). This encompasses their policies and practices related to content moderation, algorithmic decision-making, human oversight, and internal complaints-handling procedures. Such information must be communicated in clear, intelligible, and accessible language, and made publicly available in a machine-readable format. Moreover, in applying these restrictions, providers are required to act diligently, objectively, and proportionately, with due regard for the rights and legitimate interests of all parties involved—especially the fundamental rights of users, such as freedom of expression and access to information (Art. 14.3 DSA). Additionally, intermediary service providers must publish an annual report on the moderation activities carried out in the reporting period (Art. 15 DSA). Such reports shall include detailed information on the number of orders received from national authorities, notifications made by users regarding illegal or harmful content, content moderation carried out on their own initiative, use of automated means for content moderation purposes and complaints received through their internal complaints management system.

A second layer of due diligence obligations is directed at hosting service providers. They are required to establish notice-and-action mechanisms that enable individuals to report content they consider illegal or in breach of the terms and conditions, allowing the provider to assess the notice and, where appropriate, remove or disable access to the content. (Art. 16 DSA). They are also required to provide clear and specific statement of reasons to any user affected by a restriction adopted on the basis that the content was illegal or incompatible with the terms and conditions (Art. 17 DSA). In addition, hosting service providers shall promptly inform the competent national police or judicial authorities if they become aware of any information that gives rise to suspicion of the commission of a crime involving a threat to the life or safety of persons (Art. 18 DSA), which may also be considered a form of moderation (Goldman, 2021).

Additional obligations specifically target online platforms, a new category of intermediary service providers that the DSA defines as hosting service providers, who in addition to storing content at the request of a recipient of the service make that content available to the public (Art. 3.1.i DSA). This category includes social networks, online content-sharing services, e-marketplaces, app stores, or platforms for managing accommodation or transportation.

Online platforms are required to have an internal redress mechanism to deal with complaints against any decision taken on the grounds that the information provided by the recipients of the service was illegal or incompatible with the terms and conditions of the platform (Art. 20 DSA). Such complaints should be dealt with in a timely manner and in a non-discriminatory, diligent and non-arbitrary manner, although complaints submitted by 'trusted flaggers' previously recognised as such by national digital service coordinators are expected to be prioritised (Art. 22 DSA). Decisions must be made under the supervision of appropriately qualified personnel and not solely by automated means. If a user's complaint is found to be well-founded, the initial decision must be promptly reversed. Moreover, decisions taken by online platforms—including those issued through internal complaint mechanisms—are subject to review by a certified out-of-court dispute settlement body, without prejudice to the right to seek judicial remedy before a competent court (Art. 21 DSA).

As the Commission itself noted in its 2023 Initiative on Virtual Worlds, the EU has a robust, future-oriented legislative framework that already applies to several aspects of the development of virtual worlds and Web 4.0. Indeed, platforms operating different virtual worlds or proto-metaverses can readily be considered hosting service providers, and particularly online platforms, given their role in storing and disseminating information to the public (López-Tarruella and Rodríguez de las Heras, 2024). Consequently, they are subject to all the content moderation obligations previously mentioned. Nevertheless, it cannot be overlooked that the provisions of the DSA, including those that affect content moderation, were drafted primarily with Web 2.0 platforms in mind. Applying them to the metaverse is not straightforward. Key terms like 'content,' 'hosting,' and 'user-generated information' must be reinterpreted to include actions and interactions in 3D environments.

Furthermore, the DSA's requirements regarding transparency and appeal mechanisms presuppose a degree of persistence and retrievability of digital interactions. However, many interactions in the metaverse are inherently ephemeral. In cases where a user experiences harassment during an unrecorded virtual meeting, it may be difficult—if not impossible—to substantiate the violation or effectively appeal a moderation decision. The DSA's procedural guarantees thus need technological reinforcement—e.g., optional recording features, timestamped logs, or real-time reporting tools.

Decentralized virtual worlds introduce a further layer of complexity. Attributing responsibility to the platform for failing to comply with the obligations imposed by the DSA is unproblematic in the case of centralized platforms developed and controlled by a single company (e.g. Horizon Worlds

operated by Meta). However, identifying a responsible legal entity is not straightforward in metaverses built on decentralized structures, which are specifically designed to be governed not by a single organization, but by a community of users (Mienert, 2021). This is the case, for example, of Decentraland, a blockchain-based virtual reality platform that defines itself as a Decentralised Autonomous Organization (DAO). The DAO made up of the owners of the different plots, is the one that, through a democratic internal voting process, adopts all the decisions that affect the operation of Decentraland and its content moderation policy.

Conclusion

The emergence of the metaverse presents challenges to existing content moderation frameworks. While the European Union's Digital Services Act offers a solid starting point, its implementation in immersive environments is neither straightforward nor comprehensive.

The DSA's assumptions of persistent, retrievable, and text-based content often clash with the transient and spatial nature of interactions in virtual worlds. Effective moderation in these environments thus requires not only a reinterpretation of legal categories such as 'content' or 'hosting', but also the development of new technical tools and preventive mechanisms tailored to immersive settings. Furthermore, decentralized platforms introduce structural obstacles to enforcement, calling into question the DSA's capacity to ensure accountability when no central legal entity exists.

As the metaverse evolves, regulatory strategies will need to adapt accordingly, finding a workable balance between innovation and user freedom on the one hand, and safety, accountability, and the protection of fundamental rights on the other. In the end, ensuring that virtual spaces remain safe and fair will require ongoing coordination between legal frameworks, technological design, and the governance practices of the platforms themselves.

References

Ball, M. *The Metaverse – And how It will revolutionize everything*, Liveright Publishing Corp, 2022.

Castro, D., 'Content moderation in multi-user immersive experiences: AR/VR and the future of online speech', Information Technology & Innovation Foundation, 2022, https://www2.itif.org/2022-immersive-content-moderation.pdf

CJEU (Grand Chamber) of 24 September 2019, Case C-507/17 (Google), ECLI:EU:C:2019:772.

European Commission, 'An EU initiative on Web 4.0 and virtual worlds: a head start in the next technological', COM(2023)442, 2023. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13757-Virtual-worlds-metaverses-a-vision-for-openness-safety-and-respect_en

European Union, 'Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)', *Official Journal of the European Union*, L 277/1, 2022, ELI: http://data.europa.eu/eli/reg/2022/2065/oj.

Freeman, G., Samaneh Zamanifard, S., Maloney, D. and Adkins, A., 'My body, my avatar: How people perceive their avatars in social virtual reality', *CHI EA '20: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, ACM Digital Library, https://doi.org/10.1145/3334480.3382923

Goldman, E., 'Content Moderation Remedies', *Michigan Technology Law Review*, Vol. 28, 2021, pp. 1–59, https://heinonline.org/HOL/Page?handle=hein.journals/mttlr28&div=5&g_sent=1&casa_token=&collection=journals

Hine, E., 'Content moderation in the metaverse could be a new frontier to attack freedom of expression', *Philosophy and Technology*, Vol. 36, Issue 3, 2023, https://ssrn.com/abstract=4458433

Janssen M., 'Governing the metaverse', in: Dwivedia, Y. K. et al., 'Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy', *International Journal of Information Management*, Vol. 66, 2022, pp. 12–14, https://doi.org/10.1016/j.ijinfomgt.2022.102542

Jurecic, Q. and Rozenshtein, A. Z., 'Mark Zuckerberg's metaverse unlocks a new world of content moderation chaos', Lawfare, 2021, https://www.lawfareblog.com/mark-zuckerbergs-metaverse-unlocks-new-world-content-moderation-chaos

Klonick, K., 'The new governors: The people, rules, and processes governing online speech', *Harvard Law Review*, Vol. 131, No. 6, 2018, pp. 1598–1670.

López-Tarruella, A. and Rodríguez de las Heras, T., 'A European regulatory framework for the Metaverse', MetaverseUA Research Paper #1, /2024), 2024, https://ssrn.com/abstract=5024023 or http://dx.doi.org/10.2139/ssrn.5024023

Mienert, B., 'How can a decentralized autonomous organization (DAO) be legally structured?', *Legal Revolutionary Journal LRZ*, 2021, http://dx.doi.org/10.2139/ssrn.3992329

Rodríguez de las Heras Ballell, T., 'La fórmula de la DSA para resolver el "dilema de la responsabilidad de las plataformas": un equilibrio entre continuidad e innovación', in: Hernández Sainz, E., Mate Satué, L. C. and Alonso Pérez, M. T. (dir.), *La responsabilidad civil por servicios de intermediación prestados por plataformas*, Colex, 2023, pp. 25–50.

Sabri, N., Chen, B., Teoh, A., Dow, S. P., Vaccaro, K. and Elsherief, M., 'Challenges of moderating social virtual reality', *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Article No.: 384, 2023, https://doi.org/10.1145/3544548.3581329

Schulenberg, K., Li, L., Freeman, G., Zamanifard, S. and Mcneese, N., 'Towards leveraging AI-based moderation to address emergent harassment in social virtual reality', *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Article No.: 514, 2023, https://doi.org/10.1145/3544548.3581090

Singh, S., 'Everything in moderation. An analysis of how Internet platforms are using artificial intelligence to moderate user-generated content', New America's Open Technology Institute, 2019, https://d1y8sb8igg2f8e.cloudfront.net/documents/Everything_in_Moderation_2019-07-15_142127_tq36vr4.pdf

Wiederhold, B. K., 'Sexual harassment in the Metaverse', *Cyberpsychology, Behaviour, and Social Networking*, Vol, 25, Issue 8, 2022, https://doi.org/10.1089/cyber.2022.29253.editorial

5.3. Toxicity in Gaming and Virtual Environments: User Perspectives and Needs

Ouassima Belmoussi

Offlimits / Gaming content creator

Abstract

In this short report, I discuss the findings of my research on online toxicity, alongside a critical reflection on my personal experiences as a female gamer, a person of colour, and a content creator. The research explores online toxicity in gaming, fuelled by anonymity and normalisation of harmful behaviour like harassment and hate speech. Marginalised groups —especially those who identify as women, LGBTQI+, and/or people of colour— face the worst impact, leading to negative effects on their mental health and community exclusion. Based on twenty in-depth interviews, the study shows that gamers want better reporting tools, stronger moderation, more inclusivity, role models, and support systems. Tackling toxicity requires a cultural shift: bystanders must intervene, and all stakeholders —from gamers to developers and policymakers— must act. Ultimately, solutions must involve gamers in creating positive and inclusive game communities. Throughout my gaming career, I have frequently encountered sexism, hate speech, and various forms of harassment. I have repeatedly been told that gaming is not meant for women. Many women I have known have left the gaming community due to persistent sexism and hostile environments. These issues must be taken seriously: not only because of the emotional toll they take, but also because of the ways they exclude people from a space that can otherwise offer joy, community, and creativity.

Highlights

- Marginalised gamers are often the target of toxic behaviours, based on how others perceive or assume their identity, particularly in terms of gender, sexuality, race, or ethnicity, and on how these aspects often intersect.
- Many players perceive moderation tools as ineffective, which contributes to frustration among gamers and discourages reporting toxic behaviours.
- The fact that unmoderated toxic behaviours are widespread leads many players to normalize them and perceive them as acceptable, fuelling a vicious cycle in this regard.
- The presence of positive role models —particularly high-profile gamers who reject toxic behaviours and promote inclusive and respectful gaming— can help break this vicious cycle.

Introduction

My name is Ouassima, and I'm a policy officer for gaming at Helpwanted in the Netherlands. Outside of my work, I'm a content creator, livestreamer on the platforms Twitch and YouTube, and game influencer. Gaming has been a part of my life since childhood, and I have been playing multiplayer games since my teenage years. Over the years, I have had the privilege of experiencing

many positive aspects of the gaming community, but unfortunately, I have also faced challenges that have impacted my mental health.

Toxicity in gaming refers to actions like hate speech, harassment, and cyberbullying, which can be especially harsh for marginalised groups, including those who identify as women, LGBTQI+, and/or people of colour. As a female gamer of colour, I have experienced racism, sexism, and sexual harassment, often rooted in harmful stereotypes. This is common in multiplayer games, where some players believe women don't belong in gaming or can't compete at a high level.

In my community, many gamers from marginalised groups share similar experiences of online toxicity. These stories are often distressing and have led some to step away from gaming. Research supports this, showing that marginalised gamers are more likely to face toxic behaviour, often from young (white) males (Gray, 2012; Buckels et al., 2014; Cook et al., 2018; Lemercier-Dugarin et al., 2021). Despite this, many in the gaming community view toxicity as part of the experience and expect those affected to just tolerate it. I believe it is crucial to speak out. Online gaming should be fun and safe for everyone, not just a small group.

Nearly one-third of the world's population—around three billion people—are gamers according to Statista (Clement, 2025). Despite this massive number, the general public knows little about the gaming world or the abuse that millions of people face daily. Policymakers and governments often neglect the gaming industry when crafting new policies for online platforms, even though gaming is a multi-billion-dollar sector. Conversations among parents and educators tend to focus on screen time rather than the behaviour that young people face while gaming.

It is time for a culture change. Gaming has the power to connect people, develop cognitive skills like quick thinking, and much more. This is why I'm so passionate about this issue. I believe that game influencers, who have the power to shape opinions and influence a wide audience, should talk more openly about toxicity, raise awareness, and contribute to lasting change in the community. In my role as a policy officer at Helpwanted, I aim to drive this change from a policy perspective, working to create a safer, more inclusive gaming culture for all.

Research paper at Offlimits/Helpwanted

Helpwanted is an initiative of Dutch NGO Offlimits, the center of expertise for online abuse. Helpwanted consists of a helpline that helps all victims of online transgressive behaviour. By online abuse we mean the misuse of personal images and/or personal data, grooming, sextortion, cyberbullying, online toxicity in games, cyber-stalking, account hacking, fake profiles and online scam and fraud on the internet. In my capacity as policy officer, I was asked to conduct preliminary research to inventorise the kind of abuse that takes place in the online gaming world, and to assess gamers' needs after experiencing toxicity. My own experiences as a gamer formed the starting point for this research and paired with the academic expertise of my colleague Kira Esparbé Gasca, the extensive research paper "Game Over: Gamer needs in a toxic online landscape" was published in the summer of 2024.

Summary of the research paper

This publication examines online toxicity in gaming communities, where behaviours such as verbal harassment, discrimination, and cheating are aimed at harming other players in multiplayer games. The online disinhibition effect fuels this toxicity, giving individuals a sense of anonymity and detachment (Huijstee et al., 2021; Reid et al., 2022). This leads to a vicious cycle where toxic

behaviour becomes normalised, influencing others to act the same (Kowert & Cook, 2022). As toxicity grows, it often goes unrecognized, and bystanders fail to intervene (Beres et al., 2021), making it harder to break the cycle and foster a positive environment (Frommel et al., 2023; Reid et al., 2022).

Marginalised groups are most affected by this toxicity, which can lead to anxiety, stress, isolation, and depression. In the broader gaming community, it may result in reduced player retention and performance in eSports, with potential financial losses. Gamers cope with toxicity through methods such as ignoring, avoiding, reporting, or counterspeech, depending on their needs.

Based on twenty in-depth interviews, the study highlights gamers' need for better reporting features, improved moderation, more inclusivity, constructive dialogues with toxic players, positive role models, and support systems like helplines. Tackling toxicity requires a cultural shift where bystanders actively intervene and encourage positive behaviour. Gamers, platforms, parents, educators, gaming influencers, and developers all have roles to play. Game developers should create safer, more inclusive environments, and policymakers should enforce stricter guidelines, potentially through the Digital Services Act (DSA). The toxic cycle must be broken without burdening victims, and gamers should be actively contributing to the process of finding and executing a solution.

Personal observations and opinions on this study as an expert

Online toxicity targets

In this short report I have called multiple times for a cultural shift. But before we can go to recommendations on how to achieve this shift, I would like to zoom in on the culture we're planning to leave behind.

From the qualitative interview analysis, I have observed that marginalised gamers who have experienced in-game harassment are often targeted not because of their identity itself, but because of the assumptions, stereotypes, or biases others project onto them based on markers of that identity. The way someone presents themselves in the game – their voice in the voicechat, their gamertag, or the character they choose – can "give them away" and draw a target on their back. A female or feminine-sounding voice, or language that is perceived to be associated with LGBTQI+ communities or communities of colour, a gamertag with the word "girl" in it: these cues do not objectively define a person, but they activate ingrained ideas about gender, race or sexuality. It is not your skill or gameplay that determines how you are treated —it is how others perceive and interpret your (assumed) identity through their own biased lens.

I have experienced this first-hand. I have had moments throughout my gaming career where I was contemplating changing my gamertag. It has the word "girl" in it, and I have experienced extensive sexual harassment because of it. At times I have had sexist players on my own team harassing me four matches in a row, shouting slurs and abuse at me even though I had not even used the in-game voicechat. This already happened in pre-lobby, when players pick a character and the match hasn't started yet. I just wanted to have fun and level up. I had my moments where I thought changing my gamertag into something ungendered would solve the issue. It is not fun to experience sexual harassment, and it affected my mental health greatly. In the end I decided to keep my gamertag because it is a part of my online identity and I am afraid that by changing it, I will lose a part of myself. I also didn't want to give my harassers the satisfaction of "giving in". But it still hurts.

From an intersectional perspective, it is important to acknowledge that an accumulative effect can occur when aspects of identity such as gender overlap with categories such as ethnicity, skin colour, and sexual identity (Gray 2012; Gray and Leonard 2018). Players who are female, black and lesbian, for example, face a higher risk of being targeted by (various forms of) online toxicity.

For male-presenting gamers who are presumed to be white, cisgender and heterosexual, harassment is often skill-based. These players are generally perceived as neutral until the in-group has assessed their abilities. They are judged on how well or poorly they play —not on who they are. Of course, men can still fall victim to toxicity, such as bullying, hate speech, DDoS attacks or flaming. But the threat of targeted harassment is not as immediate or identity-based as it is for marginalised gamers whose voices, names or avatars mark them as 'other' from the outset.

Moderation in games and game platforms

Secondly, I have observed, that a lot of the gamers think reporting toxic users is useless because there are no noticeable consequences. They still encounter the same toxic player in their matches even if they have reported them before. This discourages the use of this functionality —if even available. All games have their own way of reporting players in-game and other tools to shield players against toxicity like muting players or blocking. Blocking works differently in every game. When you block a player on the PlayStation Network, you're shielded from receiving their messages but you still can encounter them in a game, which is absurd.

The unwillingness to report is something I see in my community as well. Some of my viewers were shocked when I reported someone on Apex Legends when he said a racist slur. They told me reporting is useless because of their own bad experiences. Some games do it better than other games, but there is still a long way to go. In the past I have had to use my social media platforms to expose toxic players who went way too far because there was no reporting system available in the game I played. My tweet went viral and reached the community manager of the game. Only then consequences against the player were taken. Why does a gamer need a big following to be taken seriously? It should be the game studio's responsibility to protect gamers by giving them a proper reporting system that works. If a game does not have a reporting system on the PlayStation console, there is no alternative to report the player through the game platform. Additionally, it is worth mentioning, that although blocking and muting provide options for players to not engage with toxicity, they do not actually constitute a long-term solution to it if toxic players are not held responsible for their behaviour. The question here is simple: When will gaming platforms finally be held responsible, and provide proper tooling to protect gamers and create a better environment for them?

Coping strategies

The coping strategies most players use, veer on the passive side. Choosing an active approach like counterspeech, requires a lot of energy and effort from the victim, and this is not always in reach. The gamers I have interviewed expressed that they are generally unwilling to respond to toxic behaviour and just mute the players. Some even said that if they choose an active approach and talk back, it would only escalate the conflict, which in their eyes was not worth it. This is consistent with the findings of other researchers in this field. Choosing an active coping mechanism is not common practice in online games, nor is seeking help after experiencing online toxicity (Reid et al.,

2022). Research by Cary and colleagues (2020) shows that only eighteen to twenty per cent of gamers take action against online toxicity.

I can empathize with this sentiment, because I know from my own experience how hard choosing an active approach can be. For me the choosing a coping mechanism depends on my mood. I usually ignore toxicity and mute it if it is possible in-game. I also often opt for the active approach by reporting the players. According to Unity Technologies (2023), 34 per cent of all gamers use a reporting tool, although the literature shows conflicting results. If I'm up for the challenge I choose to apply counterspeech, talk back and ask why they are behaving like that. Even though researchers agree that confronting toxic players with their behaviour in online games is one of the most effective ways to extinguish the behaviour (Kowert, 2020), it is not without risk. At times it can be a positive experience: some players apologized to me and said they regret their toxic behaviour. But I have also had situations where the opposite happened, and the situation only escalated. Whenever this happened, it took a toll on my mental health and made me anxious to apply this technique again, because I feared I would have to hear all the bad words again.

Vicious circle

The normalisation of toxicity in gaming communities follows a vicious circle. The toxic behaviour of some players affects others —both the direct targets and those who witness it. This exposure can cause players to mirror that behaviour, leading them to act out toxically themselves —especially if they see it going unchallenged. As a result, certain forms of negativity become perceived as 'just part of the game' and it becomes normalised. This normalisation lowers the threshold for further toxic behaviour, as players feel less inhibited when they see others doing the same (Kowert & Cook, 2022). Bystanders become less likely to intervene, and the toxicity is no longer recognised as abnormal or unacceptable (Beres et al., 2021). This dynamic reinforces itself over time, making it increasingly difficult to disrupt. Several gamers I interviewed described being shocked by the toxicity when they first started playing —but also noted that they quickly grew desensitized to it. That is precisely how deep this cycle runs.

However, this vicious circle can be broken. In our extensive research paper, we outline clear and actionable recommendations for both governments and the gaming industry to create safer, more inclusive digital environments.

Due to limitations in space for this short report, I will not go into all recommendations in detail. I have mentioned the need for a constructive legal framework above, but for the full list, please see the complete research paper[28]. However, I would like to zoom in on the recommendation where I personally feel I can contribute most, given my background.

Role models in gaming

From the interviews I have conducted for the research paper, one clear need was voiced by all participants: the need for positive role models in the community. Role models —such as gaming content creators and livestreamers— are crucial in mitigating toxicity. Gamer culture is notoriously

[28] https://offlimits.nl/assets/downloadable_files/onderzoek-game-over-offlimits-english.pdf

difficult to address, especially when a call for change comes from an "outside" perspective. Positive role models are vital partners in permeating gaming culture and creating sustainable change.

Streamers can have a massive following and have a huge influence on their —often young and impressionable— audience. It must be mentioned that a lot of streamer content nowadays is centred around toxicity, as extreme content attracts more views, and is therefore more profitable for the creator. This feeds into the cycle of toxicity, where especially the younger viewers are constantly fed with toxic content. If their favourite streamer is insulting their teammates and opponents all the time, the audience will think this behaviour is normal and funny and they will do the same when they are playing games. However, there are ways to decrease their reach: Riot Games decided to ban an exceptionally toxic content creator for verbal abuse and player harassment, which meant that he lost part of his income. Actions like these will halt the financial incentive for toxic content. Fortunately, there are also positive streamers who focus on good vibes and having fun. Game studios can play a bigger role here by encouraging and highlighting these streamers more often, for example in their games, social media or events.

As a streamer myself, I often feel the responsibility to always stay positive and kind because I know this is the most effective way to deal with it. Even if I get insulted, trashtalked, harassed, teabagged, I try to be nice because I do not want to lower myself to their level and I often think to myself that they might just have a bad day. I have noticed that when I engage with other players, it can help them realize that their behaviour is unacceptable. A lot of them have shared that they feel pressure to act toxic because that's what they have often experienced from others. It is all part of this vicious circle mentioned above. I try to break the cycle by saying things like "good game" or "you are a really good player," and it is amazing to see how many of them are genuinely shocked. They end up apologizing and admitting they regret their actions, and it is like a lightbulb moment for them. They become more aware of their toxic behaviour and try to be kinder to others. Unfortunately, this does not always extend to players who were toxic toward them. Still, I have had a lot of my viewers tell me how much they appreciate the way I handle toxic players. They have said I am a role model for them, and that I have inspired them to give online gaming another shot. It is especially rewarding when younger players or even adults share that I am the reason they have become less toxic themselves, realizing that it is not worth it and that being kind really does improve the experience for everyone. They have even built stronger friendships in the community because of it.

There is still a long way to go for the gaming community to become a safer and more inclusive online space. Regulation for key actors, like platforms and gaming studios is crucial. And although the presence of toxic behaviour and content is still widespread, there are ways to shift this by uplifting streamers and influencers who foster good vibes and inclusivity. By promoting positive role models and holding key actors accountable, we can pave the way for a gaming community that is safer, more inclusive, and welcoming to all.

References

Beres, N. A., Frommel, J., Reid, E., Mandryk, R. L., and Klarkowski, M., 'Don't you know that you're toxic: normalization of toxicity in online gaming', *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems,* 2021, pp. 1-15. https://dl.acm.org/doi/10.1145/3411764.3445157

Buckels, E. E., Trapnell, P. D., and Paulhus, D. L., 'Trolls just want to have fun', *Personality and Individual Differences*, Vol. 67, 2014, pp. 97-102, https://doi.org/10.1016/j.paid.2014.01.016

Cary, L. A., Axt, J., and Chasteen, A. L., 'The interplay of individual differences, norms, and group identification in predicting prejudiced behaviour in online video game interactions', Journal of Applied Social Psychology, Vol., 50., Issue 11, 2020, pp. 1-15, https://doi.org/10.1111/jasp.12700

Clement, J., 'Number of video game users worldwide from 2019 to 2029', Statista, 4 June 2025, https://www.statista.com/forecasts/748044/number-video-gamers-world/

Cook, C., Schaafsma, J., & Antheunis, M., 'Under the bridge: an in-depth examination of online trolling in the gaming context', *New Media & Society*, Vol. 20, Issue 9, 2018 pp. 3323-3340, https://doi.org/10.1177/1461444817748578

Gray, K. L., 'Intersecting oppressions and online communities: Examining the experiences of women of colour in Xbox Live', *Information, Communication & Society*, Vol. 15, Issue 3, pp. 411-428. https://doi.org/10.1080/1369118X.2011.642401

Gray, K. L., & Leonard, D. J., 'Woke Gaming: Digital Challenges to Oppression and Social Injustice', Seattle, WA, University of Washington Pres.

Kowert, R., 'Dark participation in games', *Frontiers in Psychology*, Vol. 11, 598947, 2020, https://doi.org/10.3389/fpsyg.2020.598947

Kowert, R., and Cook, C., 'The toxicity of our (virtual) cities: Prevalence of dark participation in games and perceived effectiveness of reporting tools', Scholarspace, 2022, https://hdl.handle.net/10125/79724

Lemercier-Dugarin, M., Romo, L., Tijus, C., & Zerhouni, O., 'Who are the Cyka Blyat?" How empathy, impulsivity, and motivations to play predict aggressive behaviors in multiplayer online games', *Cyberpsychology, Behavior, and Social Networking*, Vol. 24, Issue 1, 2021, pp. 63-69, https://doi.org/10.1089/cyber.2020.0041

Huijstee, M. van, W. Nieuwenhuizen, M. Sanders, E. Masson and P. van Boheemen, 'Online ontspoord – Een verkenning van schadelijk en immoreel gedrag op het internet Nederland' [Online derailed - An exploration of harmful and immoral behaviour on the internet Netherlands], Rathenau Institute, The Hague.

Unity Technologies, '2023 Toxicity in Multiplayer Games Report', 2023, https://create.unity.com/toxicity-report

Wijkstra, M., Rogers, K., Mandryk, R. L., Veltkamp, R. C., and Frommel, J., 'Help, my game Is toxic! First insights from a systematic literature review on intervention systems for toxic behaviors in online video games', In *Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play* pp. 3-9. https://doi.org/10.1145/3573382.3616068

6. Conclusions

6.1. Key Takeaways and Next Steps

The impact of information and communication technologies on users' well-being —and, particularly that of children and young people— is generating increasingly intense public debate. In response, research on this issue has grown exponentially in recent years (Azizan, 2024; Jun et al., 2025) and various governments have begun to propose or adopt regulatory measures aimed at minimizing potential harmful effects on the population (e.g., in Australia, Spain, the United States, China, and South Korea). However, despite the surge in research and several efforts to build scientific consensus on the matter (e.g., Capraro et al., 2025) such consensus still appears distant (Stokel-Walker, 2025), with multiple voices warning of the risks of regulating without a robust scientific foundation (Nogrady, 2024; Schneiders & Gilbert, 2024). Although discussions about the impact of digital media on well-being often focus on widely used social media platforms (e.g., Facebook or Instagram), it is important not to overlook applications in the domain of virtual worlds, which currently attract millions of users[29] and may also have a substantial impact (e.g., Frazer, 2025; Metz, 2022). Along these lines, citizens and policymakers have expressed the need to understand this impact through dedicated research (e.g., European Citizens' Panel on Virtual Worlds, 2023; European Commission, 2023).

To carry out this research as effectively as possible, this report has aimed to lay the groundwork for future research by providing and overview of existing knowledge and identifying the key issues and the most appropriate methodological approaches to address them. The dialogue we have established with experts, through their contributions to this report, as well as with the research community attending the online seminar series "Virtual worlds and Wellbeing: Setting the Research Agenda", has enabled us to outline a set of principles that could serve as a foundation for such a research agenda, which we detail below.

First, it is clear that the impact of virtual world use on users is both **multidimensional** and **multidirectional**. Research efforts in this area must avoid overly broad approaches that overlook the complexity of the phenomenon. In particular, "dose–response" frameworks that focus solely on screen time or frequency of use of a given application are unlikely to yield meaningful insights for informing public policy. Addressing this issue requires, first and foremost, considering the **characteristics of virtual world technologies**, the content they give access to, and the actions they enable (or even, *script*). These factors may influence outcomes at different levels. For example, in the case of virtual reality (VR) platforms, immersive technological features (i.e., wide field of view, spatialized sound, body tracking, etc.) enhance the feeling of presence ("being there") and can lead to a variety of outcomes —both positive and negative: from experiences of "depersonalization," to making social interactions feel more intimate and supportive, to intensifying negative feelings during episodes of harassment. It is therefore essential to examine technological features and their various use contexts in a nuanced and integrated way. Beyond the immersive properties of VR, other factors such as the presence of inappropriate content (and the incentives that creators may have to produce such content; see the contributions by Gui and Kou in this report), features with addictive potential (e.g., loot boxes), and moderation options (e.g., the ability to block users with

[29] https://prioridata.com/data/roblox-users/

inappropriate behavior) are also likely to play a key role in shaping how virtual world experiences affect user well-being.

Another key aspect to consider is the **individual characteristics** of users and the extent to which these may act as **risk** or **resilience factors**, since it is apparent that different users may be affected in different ways by the same technology. **Age** is the most obvious factor, but there are others that must be taken into consideration. For instance, in this regard, research shows that **women** and other traditionally disadvantaged groups (e.g., **ethnic and racial minorities, LGTBIQ+ users, users with disabilities**) may be at greater risk of experiencing harassment. Education, and parental guidance in the case of younger users, are also relevant factors. In turn, **users' prior experience** with the technology may be a key factor, not only as a moderator of its effects, but also (due to habituation phenomena) in influencing researchers' ability to detect those effects. Additionally, user characteristics may be central to understanding not only the negative, but also the **positive impacts** of technology. For example, virtual worlds may help some individuals from marginalized communities connect with like-minded people they might not have access to in their physical environments.

Beyond overly broad judgments about whether virtual worlds are inherently "good" or "bad," there is a clear need for research adopting a fine-grained approach to the risks and benefits of these environments for user well-being. The experts that have taken part in this report have identified several potential benefits (in terms of **social capital** and **social support**, individual **identity expression and development**, **visibility** of marginalized communities, among others), as well as risks (from **harassment** to **problematic use** and **addiction**; exposure to **inappropriate content**; **financial harm**; momentary symptoms of **depersonalization**; and **body image** issues). However, a first step toward a detailed analysis of the factors described above is to identify (and prioritize) the actual uses of these platforms and the situations users are exposed to. To do this, it will be essential to include **observational** (e.g., ethnographic) and **qualitative studies** that engage the **full range of stakeholders**, that is, not only researchers and experts, but also users (including adults, minors, and their parents, depending on the context), as well as experience designers, industry actors, and the policymakers who can ultimately integrate scientific evidence in their work.

Some of the contributing experts have also stressed the need to refine the causal theoretical frameworks and methodologies employed, in order to establish causal relationships between technology use and its impact on well-being. Future research would benefit from more widespread adoption of methods that can yield robust **causal evidence**, moving beyond the cross-sectional designs (cf. Cummings, 2018) that have dominated previous work. In this respect, **randomized controlled trials** are typically considered the gold standard for providing causal evidence of effects, although in this context they may not be feasible for practical or ethical reasons (see the contribution by Mansfield in this report). **Longitudinal designs** (e.g., Random Intercept Cross-Lagged Panel Models; Hamaker et al., 2015) may be a suitable alternative in some cases, but it is essential that they take into account **relevant confounders** (e.g., identified through prior qualitative work with experts and stakeholders). A clear and concise definition of the research question and of both theoretical and empirical estimands is also essential, as is careful consideration of the temporal scope of hypothesized effects and the possibility that those effects may not be linear.

When designing studies, a central issue is the **quality of the measurements** obtained. Much of the previous research on digital media and well-being has relied on user self-reports (e.g., in terms of screen time or online activities), which are often unreliable. Regarding the use of **platform**

usage data, companies possess a wealth of information that could help address this issue. It is therefore essential to facilitate researchers' access to these data (while always safeguarding user's privacy and other rights). Various mechanisms can be used to achieve this, ranging from one-off collaborations between research centers and companies, to the use of APIs, to data donation approaches. In this regard, collaboration with other stakeholders —such as policymakers who develop and implement regulatory frameworks (e.g., in line with Article 40 of the European Union's Digital Services Act)— may be key.

Furthermore, there is significant heterogeneity in the measures of well-being used across studies, which makes synthesizing evidence more difficult. Efforts toward **standardization** and **consensus among researchers** on which specific instruments might be preferred would be highly beneficial in this regard.

Importantly, research on the impact of virtual worlds on well-being should not only focus on identifying and measuring such impacts but also on finding **solutions** to the problems already identified. These solutions may include **technological tools** as well as **educational interventions** targeting users and other stakeholders (e.g., parents, content designers) to help minimize potential risks. Regarding technological solutions, a key aspect is **content moderation**. In this respect, the nature of interpersonal interactions (real-time, ephemeral) in virtual worlds poses a challenge for moderation mechanisms and makes these interactions difficult to trace. Moreover, monitoring them could threaten user privacy. On the other hand, while the legal framework in the EU already appears to cover the main risks for users in a general sense, in the context of virtual worlds (e.g., interactions on VR platforms), a higher level of specificity may be necessary due to their unique characteristics.

As for educational interventions, these should not be limited to end users alone but should also address other stakeholders. In addition to families —particularly in the case of applications used by minors— it is important to consider that in many of today's most successful virtual worlds, users often take on the role of content and experience creators, producing user-generated content (UGC). Therefore, educational interventions must also raise awareness about **responsible content creation**. Understanding the **skills and competencies** of designers, as well as designing and assessing educational interventions in this area, should also be part of a comprehensive research agenda on this topic.

Finally, some of the contributions from the consulted experts go beyond issues strictly related to research and also address how such research informs public policy. For example, one issue highlighted by some of the experts involved in this report is that new technologies are adopted by society at a much faster pace than scientists can investigate their impacts, meaning that both scientific conclusions and evidence-based regulatory responses may arrive too late. One potential solution would be to **streamline research funding processes** and to anticipate the study of potentially risky technologies even before they become widely adopted. Another option for reducing the response time of regulators to risks posed by digital media —suggested by some of our contributors— is to adjust the **level of evidence required** to implement preventive policies according to the potential severity of the hypothetical risks (i.e., requiring a lower standard of evidence to act in the face of more serious potential harms, and vice versa). In addition, to support more timely responses, a promising tool could be the use of **living systematic reviews**, which are continuously updated as new evidence becomes available and which synthesize that evidence in a meaningful way (e.g., by weighting or ranking results based on their robustness).

Some of our experts have also suggested that it is necessary to reflect on and reconsider the **role of industry** in ensuring that their technological products have a positive impact. This could include strengthening mechanisms for industry participation in research, co-developing new safe-by-design practices with researchers and policymakers, and testing those practices before bringing products to market, as well as increasing industry involvement in prevention and treatment programs for clearly identified risks. Reflection on these possible strategies must, in any case, be supported by solid research, as discussed above.

Moving Forward

The insights and contributions gathered throughout this report provide a solid foundation for advancing research on the impacts of emerging virtual worlds on user well-being. Informed by these perspectives, the VirtueS project will undertake a series of research activities, as detailed below:

- We will conduct a series of **qualitative studies**, employing participant observation and in-depth interview methods, to determine how children and adolescents use virtual world platforms and to identify the potential risks and benefits associated with such use.
- In parallel, we will map the **content moderation options** available on these platforms, with particular attention to those most commonly used by minors.
- In line with the above-mentioned priority of standardizing measures across research teams, we will launch a Delphi study involving experts from academia, practitioners, and policymakers, to identify which **quantitative measures of well-being** are currently available and to assess their strengths and limitations.
- Based on the outcomes of the above activities, we will determine the most relevant and appropriate well-being related variables and examine the impact of virtual world use on these variables through a **longitudinal study** with a sample of users, taking into account individual and context aspects.
- Finally, we will carry out a study involving **content creators** in virtual worlds to investigate how they integrate user well-being protection (e.g., through privacy-by-design practices or the design of moderation options).

These actions represent only initial steps in exploring the multiple dimensions involved in analyzing the impact of virtual worlds on users. A thorough understanding of this issue will require larger-scale research initiatives with active involvement from a broad range of stakeholders. By aligning research efforts around the principles highlighted by the experts in this report, we can move towards a more coherent understanding of the opportunities and risks associated with virtual worlds and support the development of safer, more empowering digital environments for all.

References

Azizan, A., 'Exploring the role of social media in mental health research: A bibliometric and content analysis' *Journal of Scientometric Research*, Vol. 13, 2024, pp. 01–08. https://doi.org/10.5530/jscires.13.1.1

Capraro, V., Globig, L., Rausch, Z., Rathje, S., Wormley, A., Olson, J., Ross, R., Aşçı, S., Bouguettaya, A., Burnell, K., Choukas-Bradley, S., Fardouly, J., Kowert, R., Lopez, R., Maheux, A., Mirea, D.-M., Ozimek, P., Selterman, D., Thiagarajan, T., … Van Bavel, J. J., *A Consensus Statement on Potential Negative Impacts of Smartphone and Social Media Use on Adolescent Mental Health* (SSRN Scholarly Paper 5256747), Social Science Research Network, 2025. https://doi.org/10.2139/ssrn.5256747

Cummings, C. L., 'Cross-sectional design', *The SAGE Encyclopedia of Communication Research Methods.* Thousand Oaks: SAGE Publications Inc., 2018.

Dimitra, G., Lee, Y. H., MacDonald, M., Catton, A. M., Penbegullu, Z. K. K., and Pulido Lock, J. A.. *IP and Metaverse(s) – an externally commissioned research report*. GOV.UK., 2024, https://www.gov.uk/government/publications/ip-and-metaverses-an-externally-commissioned-research-report/ip-and-metaverses-an-externally-commissioned-research-report

Downey, S., 'History of the (virtual) worlds', *The Journal of Technology Studies*, Vol. 40, Issue 2, 2014, 54–66. https://doi.org/10.21061/jots.v40i2.a.1

European Citizens' Panel on Virtual Worlds, *Panel output*, 2023. https://citizens.ec.europa.eu/european-citizens-panels/virtual-worlds-panel_en

European Commission, *An EU initiative on Web 4.0 and virtual worlds: A head start in the next technological transition*, 2023.

Frazer, G., 'Honest or unrealistic? Roblox boss's online safety advice sparks debate', *BBC,* 2025. https://www.bbc.com/news/articles/clynwgdrr48o

Hamaker, E. L., Kuiper, R. M., and Grasman, R. P. P. P., 'A critique of the cross-lagged panel model.' *Psychological Methods, Vol. 20*, Issue 1, 2015, pp. 102–116, https://doi.org/10.1037/a0038889

Hupont Torres, I., Charisi, V., de Prato, G., Pogorzelska, K., Schade, S., Kotsev, A., Sobolewski, M., Duch Brown, N., Calza, E., and Dunker, C., *Next generation virtual worlds: Societal, technological, economic and policy challenges for the EU*, Joint Research Centre, 2023. https://econpapers.repec.org/paper/iptiptwpa/jrc133757.htm

Jun, G., Xu, J., Alivi, M. A., Zhewen, F., Dharejo, N., and Brony, M., 'Impacts of digital media on children's well-being: A bibliometric analysis', *Online Journal of Communication and Media Technologies*, Vol. 15, Issue 1, e202501, 2025. https://doi.org/10.30935/ojcmt/15696

King, D. L., Haagsma, M. C., Delfabbro, P. H., Gradisar, M., and Griffiths, M. D., 'Toward a consensus definition of pathological video-gaming: A systematic review of psychometric assessment tools', *Clinical Psychology Review*, Vol. 33, Issue 3, 2013, 331–342. https://doi.org/10.1016/j.cpr.2013.01.002

Meier, A., and Reinecke, L., 'Computer-mediated communication, social media, and mental health: A conceptual and empirical meta-review', *Communication Research*, Vol. 48, Issue 8, 2021, 1182–1209. https://doi.org/10.1177/0093650220958224

Metz, R., 'Harassment is a problem in VR, and it's likely to get worse', *CNN Business*, 2022. https://edition.cnn.com/2022/05/05/tech/virtual-reality-harassment

Nogrady, B., 'Australia bans under-16s from social media "to protect their development."' *British Medical Journal Publishing Group*, 2024. https://www.bmj.com/content/387/bmj.q2724.full

Orben, A., 'The Sisyphean cycle of technology panics', *Perspectives on Psychological Science*, Vol. 15, Issue 5, 2020, 1143–1157. https://doi.org/10.1177/1745691620919372

Schneiders, P., and Gilbert, A., 'Banning children's social media use: A wave of symbolic regulations, but at what cost?', *Internet Policy Review*, 2024. https://policyreview.info/articles/news/banning-childrens-social-media-use/1744?fbclid=IwZXh0bgNhZW0CMTAAAR21berT2XLf6XXNDDqooeWPIcccqTK1ubiAV7h9IDwwL9LBAwWLtlTNFqw_aem_Ad4EM_EXuYyB-1ey7csp9B6tq67TDJRs-BpbHM08obb0STCP_bUiYx-OXRR3ClBESZ6JrlZ-emX4xqLspCBpqvRp

Stokel-Walker, C. ,'Attempt to reach expert consensus on teens and phones ends in argument.' *New Scientist*, 2025. https://www.newscientist.com/article/2480657-attempt-to-reach-expert-consensus-on-teens-and-phones-ends-in-argument/

List of abbreviations and definitions

Abbreviations	Definitions
AI	Artificial Intelligence
API	Application Programming Interface
AR	Augmented Reality
EDPB	European Data Protection Board
DMA	Digital Markets Act
DSA	Digital Services Act
GD	Gaming Disorder
GDPR	General Data Protection Regulation
VLOP	Very Large Online Platforms
VR	Virtual Reality
XR	eXtended Reality

List of figures

List of tables

Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
— at the following standard number: +32 22999696,
— via the following form: european-union.europa.eu/contact-eu/write-us_en.

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

EU open data

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society

Scan the QR code to visit:

The Joint Research Centre: EU Science Hub
https://joint-research-centre.ec.europa.eu

Publications Office
of the European Union