RMIT UNIVERSITY Digital Ethnography Research Centre

THE UNIVERSITY OF MELBOURNE

Digital harms: Consistency in definition, understanding and action

Concept paper

March 2026
Digital Ethnography Research Centre

Acknowledgement of Country

We acknowledge the people of the Woi wurrung and Boon wurrung language groups of the eastern Kulin Nation on whose unceded lands we conduct our research, teaching, and service. We pay respects to Ancestors and Elders past and present who have always cared for Country. We also acknowledge the Traditional Custodians and their Ancestors of the lands and waterways across Australia where we conduct our research and community engagement. We honour the work of emerging leaders and pay our respect to Country as a lifeworld that sustains all.

Authors

Rob Cover, George Buchanan, Bernardo Figueiredo, Daniel X. Harris, Nicola Henry, Edward Hurcombe, Piers Howe, David Micallef, Dana McKay, Philip Pond, Nicole Shackleton, Senuri Wijenayake, Xiangmin Zhou, Joel Humphries, Adnan Alamri, Rhyle Simcock.

Suggested citation

Cover, R., Buchanan, G., Figueiredo, B., Harris, D. X., Henry, N., Hurcombe, E., Howe, P., Micallef, D., McKay, D., Pond, P., Shackleton, N., Wijenayake, S., Zhou, X., Humphries, J., Alamri, A., & Simcock, R. (2026). *Digital harms: Consistency in definition, understanding and action* (Concept paper). RMIT University, Digital Ethnography Research Centre, DOI: 10.60836/s2w0-vn39

Artwork ‘Sentient’ by Hollie Johnson

Executive Summary

Purpose and scope

Digital harms – Including abuse, harassment, bullying, hate speech, disinformation, doxxing, image-based abuse, deepfakes, illicit promotion of harmful productions, scams and other adversarial practices – have become a defining feature of contemporary digital life. This concept paper responds to growing global concern that existing approaches to online safety are no longer adequate to protect individuals, communities or the digital environments that sustain public life. Drawing on interdisciplinary scholarship across law, media and communication studies, gender and cultural studies, psychology, information science, computer science, human–computer interaction, criminology and public health, we identify a core problem underpinning current failures: ***the absence of clear, shared and internationally intelligible definitions of digital harm.***

This collaborative, interdisciplinary paper argues that without a common taxonomy of digital harms, prevention efforts, regulation, platform governance, education, and research remain fragmented, reactive, and often ineffective. While enforcement capacity, legal constraints, and platform incentives remain major structural barriers to addressing digital harms, definitional fragmentation operates as a foundational problem that undermines coherent prevention, governance, and accountability across all of these domains. New policy instruments such as digital duties of care and safety-by-design principles represent important shifts away from post-hoc moderation, however, these initiatives are currently being layered onto a system that lacks conceptual clarity about what constitutes harm, how harm occurs, and who or what is harmed. This executive summary outlines the scale of the problem, the sources of definitional incoherence, and our central proposal: reframing digital harm as both injury to individuals (or groups of individuals) and damage to the digital ecology (and wider society) itself.

> “Our research identifies definitional fragmentation as a central barrier to effective action”

The scale and urgency of digital harms

Evidence from multiple jurisdictions demonstrates that digital harms are widespread, growing, and increasingly recognised by the public. Large proportions of adults report direct experience of online abuse, harassment, hate speech, disinformation, and other harms, with many more encountering them as bystanders. New technologies—particularly generative artificial intelligence—have expanded the range and intensity of harms through impersonation, malicious synthetic media, and automated amplification. At the same time, trust in digital platforms as spaces for information, participation, and social connection has declined.

Governments have responded with inquiries, legislation, and regulatory reform, while platforms have introduced extensive content moderation frameworks. However, there is growing recognition that moderation alone is insufficient. Content moderation faces structural constraints: scale, cost, epistemic uncertainty, legal limitations, and the difficulty of interpreting context, intent, and cumulative impact. Recent reductions in platform investment in moderation have further weakened its effectiveness. The result is a persistent gap between public expectations of safety and the lived reality of digital engagement.

Why current responses are not working

Our research identifies **definitional fragmentation** as a central barrier to effective action. Across scholarship, law and platform governance, key terms – such as hate speech, harassment, cyberbullying, abuse, trolling, disinformation and offensive content – are used inconsistently, often to describe overlapping or divergent phenomena. This fragmentation manifests in several ways:

1. **Disciplinary silos**: Different academic fields prioritise different dimensions of harm – psychological injury, legal wrongdoing, cultural toxicity or algorithmic amplification – without a shared framework that enables comparison or synthesis. Moreover, there is disciplinary divergence in acceptable solutions, with some focusing on perpetrators of harm, some on legal remedy, some on digital design, and some on victim-survivor experiences. Each approach is an important contributor to remediation and prevention of harm but systemic and social change requires they be thought together.
2. **Jurisdictional divergence**: Laws across countries apply different thresholds for harm, use different terminology, and recognise different protected categories, producing a fragmented regulatory landscape for a global, cross-platform environment. And there are, of course, countries that have no laws related to digital harms.

“Definitional complexity undermines trust and limits grassroots support for reform”

3. **Platform inconsistency**: Platforms define harms primarily through proprietary, content-focused policies that vary widely and rarely account for cumulative or ecological effects. Harm is typically inferred from individual interactions with platforms rather than patterns, behaviours or systemic amplification.
4. **Public unintelligibility**: Users struggle to understand what counts as harm, how to report it, or why certain content is removed while other harmful material persists. Definitional complexity undermines trust and limits grassroots support for reform.

Together, these inconsistencies produce what we describe as a **crisis of intelligibility**. Digital harm is widely felt and recognised, yet poorly named and inadequately conceptualised. Without a shared language, there is no stable foundation for prevention, regulation, education or accountability.

Reframing harm: individuals and the digital ecology

Our core conceptual intervention is to reframe digital harm as operating across **two interconnected domains**:

1. **Interpersonal harm**, where individuals – and groups or classes of individuals – experience injury through abuse, harassment, humiliation, misrepresentation, disinformation and other practices that undermine dignity, safety, reputation and wellbeing.
2. **Ecological harm**, where the digital environment itself is toxified by persistent hostility, manipulation, polarisation, misinformation and exploitative design, reducing trust, participation and the conditions for collective life. A digital ecology that is toxified by harms becomes, in itself, toxic for other users.

This ecological framing draws on established understandings of harm in environmental, public health and social systems. Digital platforms are not merely conduits for interaction but shared environments that shape behaviour, norms and possibilities for engagement. When these environments become toxic, harm extends beyond direct victims to bystanders, future users and democratic institutions. Fear, withdrawal and normalisation of hostility are themselves forms of harm that accumulate over time.

By foregrounding ecological harm, we offer a way out of current definitional impasses. Rather than adjudicating intent, offensiveness or legal thresholds in isolation, harm can be understood in terms of damage to the conditions that enable safe, meaningful, inclusive, democratic and civic digital participation.

> “Fear, withdrawal and normalisation of hostility are themselves forms of harm that accumulate over time”

Towards a shared taxonomy of digital harms

We argue that developing a ***multi-sector, internationally relevant taxonomy of digital harms*** is a necessary precondition for meaningful reform. Such a taxonomy would:

- Provide shared language across platforms, governments, researchers, educators and communities.
- Enable comparative research and evidence-based policy.
- Support clearer reporting and remedy for users.
- Inform safety-by-design, duty-of-care obligations and digital literacy initiatives.
- Allow emergent harms (e.g., AI-driven manipulation) to be identified early.

Rather than relying solely on content categories, the taxonomy should incorporate a matrix of harm across multiple dimensions, including injury, severity, scale, virality, intent and ecological impact. It should be flexible, adaptive and grounded in lived experience,

while recognising that harms often intersect and reinforce one another.

Recommendations

On this basis of the concept paper, we make four core initial recommendations for governments, platforms, scholars and civil society:

1. **Develop a shared global taxonomy of digital harms**
   Governments, platforms, researchers, educators, health practitioners and community advocates should collaborate – under the leadership of supranational organisations – to develop an agreed, flexible taxonomy of digital harms within five years.
2. **Ground definitions in injury and ecological harm**
   The taxonomy should prioritise demonstrated injury and harm, recognise intersections between harm types and foreground damage to the digital ecology rather than focusing on harms defined primarily by narrow, individual or isolated incidents.
3. **Use the taxonomy to drive prevention and education**
   A shared taxonomy should underpin safety-by-design, digital duties of care and digital education for children and adults, emphasising why harmful practices damage individuals, communities and the digital environment.
4. **Address the inter-platform and interjurisdictional reality of harm**
   Regulatory and platform frameworks must recognise that digital harms occur across platforms and borders. Governments and platforms should cooperate to develop safety-by-design and duty-of-care mechanisms that reflect the post-national nature of digital communication.

Definitional clarity is not just a technical exercise but as a foundational step toward a healthier digital future. By recognising digital harm as both interpersonal injury (individual or collective) and ecological degradation, it offers a coherent framework for prevention, governance and collective responsibility. A shared taxonomy of digital harms would enable more consistent regulation, more effective platform accountability and stronger public understanding – supporting a digital environment in which individuals and communities can participate safely, meaningfully and sustainably.

Contents

1.0 Introduction

This concept paper has been authored by an interdisciplinary group of scholars concerned about the serious increase in digital harms. These include abuse, harassment, disinformation, doxxing, image-based abuse, deepfakes and other malicious synthetic media, trolling and the widespread growth in adversary, enraged communication that is harming individuals and the wider digital ecology as a setting for public debate, private discourse, entertainment, learning, democratic deliberation and creative innovation. While this paper engages with a broad range of digital harms, including scams, fraud, privacy violations and algorithmic bias, its primary analytical focus is on harms that arise through, or are amplified by, digital communicative environments. We use communicative harm as an organising lens because many forms of digital harm, even when not purely interpersonal, depend on, or are enabled by, platform architectures, information flows and interactional dynamics. Where non-communicative harms are discussed, they are included insofar as they intersect with, or reshape, the digital ecology of communication.

The public globally recognises the growth of digital harms over the past decade and this awareness has been well-documented. A 2021 Pew Research Centre study found that 41% of adults in the United States have experienced some form of online abuse – a tripling of the rate in five years – and 25% of adults experienced more severe forms of online abuse, such as threats, stalking, sexual harassment and image-based abuse (Vogels, 2021; Schoenebeck, Lampe and Triêu 2023). An estimated 14% of Australian adults are subject to hate speech and substantially more to other forms of online hostility (eSafety Commissioner, 2020), with a reported 70% of Australian adults experiencing at least one digital harm during 2022 (eSafety Commissioner, 2022). 61% of European Union residents state they often come across disinformation (European Commission, 2024), while 43% of people in the United Kingdom encountered at least one troubling deepfake in the past six months to June 2024 (Ofcom, 2024).

> “Content moderation not only is not working well, but resource needs present insurmountable challenges, epistemic problems and technical constraints”

About 5% of people in Australia report that they have been doxxed (Moseley, 2024), while 31% of respondents in a multi-country study say they have personally experienced or observed “bias or unfair treatment” from AI use (KPMG and The University of Melbourne, 2025). A survey of women conducted for Amnesty International found that one-third had experienced some form of online harassment (Amnesty International, 2018). In China, a poll of more than 2,000 social media users found 40% had experienced online abuse, with 16% of victim-survivors experiencing suicidality as a result (Radio Free Asia, 2022). That is, the global public is clearly recognising – and in many cases directly experiencing – the fact that the digital environment is

toxic, unsafe, adversarial and harmful, and that the risk of being harmed is high despite the benefits of digital engagement, participatory creativity and social and workplace affordances of digital culture.

Over the past half-decade, there has been an unsurprising growth in demands for the prevention of digital harms rather than the interventional and reactive framework of moderation after harmful communication has circulated. Increasing rates of online hostility has culminated in public inquiries, law reform, and new policy initiatives in a range of jurisdictions (Flew, 2021), concerns about wellbeing and mental health (Keighley, 2022), and more stringent, cooperative state regulation of platforms (Christchurch Call, 2019). Much of this for a long time was a call for better moderation of social media and a stronger investment in moderators by platforms (Gillespie, 2018) with regulatory pressure through various legislation to require platforms to remove reported content – especially hate speech, doxxing and privacy breaches – within a certain time frame. However, it has been increasingly recognised that content moderation not only is not working well, but that the resource needs present insurmountable challenges, epistemic problems and technical constraints, whether that is moderation by human workers or automated tools (Gorwa et al., 2020; Blackwell, 2025; Tobi, 2024). These structural limits do not negate the importance of conceptual clarity; rather, they highlight how fragmented definitions of harm compound enforcement difficulties by obscuring what should be prioritised, measured, and prevented. At the same time, encouraged by the second Trump Administration in the United States, many major platforms have curtailed their investment in moderation and policing content (Noman, 2025). Government and community expectations that moderation alone is a solution to digital safety concerns are arguably now outdated.

The pressure for better, more preventative responses to digital harms has come in a range of alternative forms, including attempts to implement **duty of care** requirements on platforms. The United Kingdom's Online Safety Act 2023 embeds a duty of care on in-scope digital service providers to assess and mitigate risks of harm to users, including illegal and harmful content, and through the Digital Services Act the European Union expects platforms to undertake due diligence to mitigate risks to users. The Australian government has proposed legislating a Digital Duty of Care which would place greater responsibility on platforms for ensuring user safety from a range of harms, and other jurisdictions

including India, Brazil, South Korea, and Singapore have indicated a desire for regulatory frameworks that include proactive responsibilities for platforms.

hateful, adversarial, fraudulent, and disinformation content continues to circulate. This content continues to erode trust in credible information

> “Safety-by-design embeds harm prevention into digital platforms from the outset, rather than relying predominantly on moderation after harm occurs”

A second preventative approach is proffered in the call for **safety-by-design** principles. Safety-by-design embeds harm prevention into digital platforms from the outset, rather than relying predominantly on moderation after harm occurs. Principles include risk assessment during product development, default privacy and safety settings, transparent reporting tools, and user empowerment. Initiatives involve age-appropriate design codes, abuse-resistant interface design, algorithmic impact assessments, and ongoing monitoring of emergent harms. Together, these approaches shift responsibility toward platforms to anticipate risks, protect users, and foster healthier digital cultures through accountability and governance mechanisms.

In other words, there is widespread recognition of a key, international problem with platforms and the proliferation of digital harms, and an acknowledgment that current national and international regulatory principles, platform practices and user behaviour are not keeping users safe and free from a very high risk of encountering harmful, problematic, abusive or hateful content, attempted fraud, doxxing and other harms. Despite this recognition, and despite legislative and policy reform in many countries as well as platform policies ostensibly designed to protect users, extremist,

and significantly limits expectations of participating online without being subject to abuse and harassment, hate speech or other kinds of adversities.

We argue that new, preventative approaches are warranted to what has become a global ‘wicked problem’ over the past decade, but we argue that cobbling duties of care and safety-by-design principles into existing regulatory, interventional and cultural frameworks of digital culture is not yet workable, because there is a **lack of agreement on key definitions, terms and meanings** in the lexicon of digital harms, local and global uncertainty about the actual **injury** of harms, and an overwhelming focus on protecting users as individual platform ‘customers’ without regard for the **wider toxification of the digital ecology**. Alongside the new approaches to prevention is the need for a ***courageous alternative vision*** of digital culture – that requires broad, international community agreement. For any agreement there is a need to build ***clarity and shared understandings*** and to find nodes around which those definitional understandings of the harm of digital harms can be understood and recognised.

One key problem marking contemporary research is disciplinary silo-isation, which creates roadblocks for addressing so-

called ‘wicked problems’ on the scale of digital harm in its interjurisdictional setting. As scholarship and public concern about the toxification of digital communication has grown, different disciplinary areas have taken different approaches and terminology, which has made it difficult to develop comparative studies to understand the true scale of the problem, to evaluate the effectiveness of remedies and interventions, and to develop educational strategies that encourage more ethical engagement and cohabitation in online settings. In that respect, we have brought together scholars from a wide array of disciplines – human-computer interaction; psychology; law; computer science; media and communication studies; sociology and criminology; gender studies, marketing and business studies – to begin the process of smashing silos in favour of shared, mutually-enhancing dialogue, particularly across the STEM and HASS divide.

Secondly, a lack of clear, translatable definitions, concepts, meanings and significations among users has serious effects on the ability to gain support for change, to intervene with platforms, to develop multilateral prevention strategies with grassroots support and arguably for the functioning of society and social relations both in online and offline contexts. Some of our research has shown that definitional complexity, competing terminology and unclear concepts are not only making remedy more difficult but are actively damaging digital culture, social relations, political frameworks and social infrastructures. For example, everyday users struggle with the proliferating terminology and concepts related to digital harms and therefore are unclear on whether or not there has been a harm. Hate speech is sometimes used to describe mild insults, many users remain unfamiliar with terms at the time of reporting (Cover, Beckett et al., 2025), courts struggle with the complex legal framework and interjurisdictional terminology differences for determining penalties for serious perpetrators under the law (Shackleton, 2024; Cover, Simcock and Humphries, 2025), discourse on online safety has become so complex that there is a tendency for researchers and policy-makers to only focus on isolated forms of harm (Wijenayake et al., 2025). Different approaches to informational polarisation make it difficult to develop new theoretical and empirical remedies (Pond, 2024). Platforms have been found to use vague terminology to describe problems or to refuse definition of a problem in public announcements (Hurcombe et al., 2025), and what is meant when politicians and leaders talk of social harmony and civil society online is no longer clear (Duff, 2026). A lack of clarity on positive and negative misinformation makes it difficult to understand the impact on assessments of others (Mickelberg et al., 2024). A lack of clarity on terminology and meaning around malinformation and misinformation has been shown to sponsor ambivalence towards veracity (Bahl et al., 2025). Generational differences promote different concepts, attitudes and definitions in relation to digital harm (Aleti et al., 2025) and evidence is now showing that definitional uncertainty results in backlash to well-intended regulatory measures because the community finds the intention difficult to navigate (Shackleton, 2024). The impact of recommender systems on exacerbating harm is recognised but not well articulated (Zhou et al., 2024)

Addressing the toxicity of the digital ecology is an urgent and complex challenge. Social media and digital platforms have become central to public life, connecting billions of users across communication, work and leisure. However, harmful content, behaviours

and adversarial forms of engagement (collectively referred to as 'digital harms') remain pervasive. This increasingly hostile communication environment is corroding interpersonal relationships, distorting political discourse and electoral processes, increasing polarisation (Törnberg, 2022), and undermining everyday digital interaction, from gaming and entertainment to personal expression online. Against this backdrop, there is a growing public appetite for remedy and reform (Flew, 2021; Jane, 2015), even as many digital platforms continue to resist substantive change (Thompson & Conger, 2025).

There is a pressing need for a clear, consistent and mutually intelligible understanding of what constitutes digital harm and how it impacts individuals and communities. Despite widespread recognition of the problem, accounts of digital harm are not shared across scholarly fields, interjurisdictional legislation and platforms. This gap is underscored by three persistent challenges:

1. inconsistent definitions and standards across platforms and jurisdictions, leaving users, policymakers and courts uncertain about rights, responsibilities and expectations;
2. different emphases on whether digital harms should be understood through the lens of perpetrator behaviour and intent, textual or visual content, measurable injury to victims, or system design and algorithmic amplification; and
3. reactive rather than preventive responses, with most interventions occurring post hoc rather than through proactive strategies focused on digital literacy, citizenship, and campaigns promoting the harmonious and rational use of digital services (Suzor, 2019).

The absence of a shared, multi-sectoral and interjurisdictional taxonomy of key terminology hinders effective platform remedies, policy responses and regulatory development. Inconsistent definitions of digital harm across platforms and jurisdictional differences in how governments regulate digital environments have produced a fragmented and often conflicting vocabulary that limits coordination and weakens collective capacity to respond to harm. An important first step in addressing these issues is the development of a shared, clearly articulated, and widely recognised multi-sectoral and interjurisdictional taxonomy of digital harms.

The purpose of this paper is to provide a timely thought-piece on the challenge of defining and responding to digital harms. It highlights emerging opportunities to build a shared taxonomy of harm and to rethink how injury to users and damage to the digital ecology are understood across disciplines and jurisdictions. The proposals outlined here are not intended as a definitive framework but as a foundation for wider discussion among policymakers, platforms, researchers and communities about how we might collectively build safer, more harmonious digital environments.

2.0 Framing the crisis: Intelligibility and the need for a taxonomy

2.1 The challenge of naming and classifying "digital harm"

Efforts to explain digital harm are complicated by a basic lack of clarity: there is no standard language to define what is wrong in the digital environment. The term 'digital harm' is frequently used, but it covers a changing range of practices, including abusive language, harassment, doxxing, synthetic media, algorithmic discrimination, and organised disinformation, which are described differently across platforms, fields, and jurisdictions (Mitts, 2025; Cover, Simcock and Humphries 2025). This fragmentation is not only conceptual or a matter of differences in interpretation but exists widely in practice: different jurisdictions categorise identical behaviours differently, and different platforms use diverse and sometimes inconsistent definitions about the digital harms, resulting in incompatible taxonomies that obstruct coordinated action (Digital Action, 2023; World Economic Forum, 2023) and community understanding, including particularly among those who have been harmed (Cover, Beckett et al., 2025). But most importantly, as different platforms and jurisdictions develop or adopt their own terminology, it fails to recognise that harms occur in the interjurisdictional space, and that perpetration occurs in the cross-platform environment (Mitts, 2025). Without agreed, shared cross-jurisdictional and cross-platform clarity, we are likely to remain 'stuck' in a reactive, inconsistent digital safety framework that will remain difficult for users, policymakers, and change advocates to navigate, or in which to build genuine prevention initiatives.

Given the high rates of abuse, disinformation and other harms, the increasing perception that digital platforms are untrustworthy, that some digital use may be unhealthy or 'toxic', and the knee-jerk responses of some governments to keep users safe through restricting access (Cover, Humphries et al. 2025), we argue that the digital environment should now be described as one that is 'in crisis'. A crisis is commonly defined as a significant disruption to normal functioning that threatens essential goals, values, or stability, requiring urgent decision-making under conditions of uncertainty. A genuine crisis occurs when there are no thinkable solutions or remedies, usually because structural conditions prevent that thinking, leaving a society or community at a difficult 'cross-roads'. In terms of digital harms, crisis is invoked not because platforms are deemed a setting of risk for which there are knowable, tested and agreed solutions, but because there is an inherent and profound *unintelligibility* caused by the inconsistent language, concepts and terminology that are necessary for understanding the problem, proposing solutions, and gaining grassroots agreement for those solutions. Any remedy to this crisis must begin by developing shared agreement on digital harms terminology, the international translatability of that terminology, and

> "the digital environment should now be described as one that is 'in crisis'"

thereby building concepts and practices towards prevention across different sectors, platforms, communities, legislative initiatives, case law and digital capabilities education. A bold, new cross-sectoral framework is a key starting point for ensuring the safety of a global population of over 5 billion users by detoxifying a harmful environment.

Recognising and classifying digital harm is also becoming more difficult as new technologies create new categories of concern. The widespread introduction of generative AI has contributed new and complex vectors for impersonation, manipulation, and harassment, including the use of synthetic media to defame, defraud and mislead, with real-world examples documented in over 190 global cases of abuse (Marshall et al., 2024). At the same time, algorithmic systems are changing and making worse long-standing problems like racism, misogyny, and attacks on people's reputations in ways that make them more difficult to detect, manage, or ameliorate (Shelby et al., 2023). Recent research on abusive language categorisation demonstrates that even cutting-edge machine learning models fail because the concepts they rely on are inconsistently defined across datasets and taxonomies, leading to contradictory outputs and confusion in the field (Moghaddam et al., 2025). As a result, digital environments suffer from a growing 'descriptive deficit': the damage is visible and widely felt, but we lack a consistent vocabulary to explain what it is or how it works.

> "digital harm functions as both interpersonal harm and environmental damage"

2.2 The case for conceptual clarity and shared language

A shared conceptual and definitional understanding, an internationally agreed taxonomy of digital harms, is crucial because the stakes extend well beyond individual experiences of harm or discomfort. The cumulative impact of digital harm undermines the broader digital environment by eroding trust, distorting information flows, polarising communities,

and reducing people's ability or willingness to engage in public life. According to research on technology-facilitated abuse (Koukopoulos, Janickyj, and Tanczer, 2025), harm is seldom isolated: abusive behaviours often spread across platforms, relationships, jurisdictions and contexts, threatening not only an individual's safety but also the *integrity* of their social and informational environments. In this sense, digital harm functions as **both** *interpersonal harm and environmental damage*, necessitating a paradigm capable of recognising both.

Therefore, developing a taxonomy is not just an analytical exercise to enable more systematic and international comparative research, but a practical tool for making digital harm intelligible across the sectors who are key stakeholders in the prevention of harms, the treatment of harmful effects when they arise, the legal remediation in cases where that is necessary and the education that requires all users to understand why harming others online is wrong.

Taxonomies have been crucial in cybersecurity research for elucidating the propagation, interaction, and escalation of harms, hence facilitating the coordination of actions among actors with varying responsibilities (Agrafiotis et al., 2018). A similar challenge exists in the digital harm sphere, where definitions are fragmented and inconsistent. Attempts to create systematised definitions or 'typologies' (e.g., World Economic Forum, 2023; Enock et al., 2023; Tong, 2025) have been at best only partial, overly focused on individual nation-state regulation, or overly focused on interpersonal harms rather than acknowledging harm to the digital ecology.

Alongside the lack of consistent definition among scholars and supranational organisations, individual platforms rely on proprietary policies and moderation categories, while regulators use different statutory definitions, and civil society organisations create their own damage labels based on advocacy or community needs. The consequence is a patchwork of overlapping but incompatible terminologies that hide rather than enlighten the type and severity of harm (World Economic Forum, 2023; Digital Action, 2023). A uniform, multi-sectoral taxonomy would serve as a common reference point, improving clarity, facilitating comparability, and strengthening accountability across an increasingly cross-platform, cross-jurisdictional digital ecosystem.

Additionally, a shared taxonomy enhances public intelligibility. People do not explain their experiences in legal or technical terms; instead, they convey harm, such as feeling unsafe, being targeted, having photographs exploited, or losing trust. A taxonomy based on lived experience, behaviour, and impact enables simpler reporting processes, consistent platform responses, and more accessible public communication. It also enables emergent harms such as AI-driven manipulation, coordinated amplification, or synthetic sexual imagery to be identified and handled early on, rather than after they have caused widespread harm. A shared taxonomy serves as a foundation for prevention, safety, and healthier digital environments.

3.0 Definitional challenges and conceptual gaps

Despite growing public, regulatory and scholarly attention, there is still no shared understanding of what constitutes a digital harm or how it should be identified, measured or addressed. Scholarship, legislation and platform policies each use their own definitions and thresholds, often describing the same behaviours in incompatible ways. These inconsistencies create conceptual and operational gaps that hinder effective governance and obscure the full spectrum of injury experienced by users and the digital ecology. This section examines these definitional divergences and the challenges they pose for taxonomy-building, regulation and platform accountability.

3.1 Scholarly definitions of digital harm

Digital harms

"Digital harms" is used across scholarship and journalism as an umbrella term for harmful content, behaviours, user impacts, and system-level design features such as algorithmic amplification (Nash & Felton 2024). It has utility as a broad classificatory concept that encompasses a wide range of harms, including technology-facilitated abuse and harassment (Henry, Flynn and Powell, 2020), trolling, doxxing, deepfakes (Flynn, Powell et al., 2025), disinformation, scams, AI bias and more (Holly 2024; Kelly 2023). In doing so, it recognises that: (1) many harms share common systemic drivers, (2) perpetration often spans multiple harm types (e.g., disinformation intertwined with hate speech), and (3) digital systems, cultural practices, and platform environments can themselves generate injury through design choices and amplification dynamics.

However, definitions of digital harm vary widely across fields. Cyberpsychology prioritises trauma and distress arising from cyberbullying or algorithmic harm (Kowalski et al., 2014; Noble 2018). Media studies situate harm in systemic inequalities shaped by platforms and recommender systems that amplify disinformation and adversarial speech (Benjamin 2019; Jane 2017). Legal scholarship typically frames harm as malicious or negligent conduct (Franks 2019). Interdisciplinary research approaches digital harm more broadly, conceptualising it as a constellation of psychological, social, cultural, and technological impacts produced through digital environments (Buil-Gil et al., 2023; Kelly 2023).

Nevertheless, the term has significant value for the reasons described above. Additionally, it helps point to problematic digital content, use, systems and the cultural meanings around them have become an important and pressing social concern (Mitts, 2025) and helps to describe the fact that not only are some users unwittingly injured on digital platforms, but that the harms themselves are toxifying the digital environment in a way that repeats the risk of that harm.

Hate speech

Although there is no universally agreed-upon definition of hate speech (Matsuda et al., 2018) it is widely recognised as a serious form of digital harm – and its increased use has garnered the attention of governments concerned about the connection with social discord, instability and division. Additionally, digital hate speech has a notable reputation, also, because of the growth in obvious,

ostensible and unashamed hate speech targeting racial and ethnic minorities, which is viewed by many users as an affront to the community (Waldron, 2012; Strossen, 2018; Carlson, 2021). In Australia, 34% of adults encountered what they defined as hate speech, and 18% were targets of it (eSafety Commissioner, 2025).

Despite the seriousness of hate speech, its definition and boundaries remain contested. Many governments and supranational organisations have offered definitions, although they remain inconsistent and contested. The United Nations defines it as communication that attacks or demeans individuals or groups on the basis of protected identity characteristics such as race, religion, ethnicity, gender or nationality (United Nations, 2025). The UN's definition is grounded the by *Universal Declaration of Human Rights* (1948) which provides protections against discrimination while establishing a right to freedom of expression (although not without limits), and in the *International Covenant on Civil and Political Rights* (1966, effective from 1976 with its thirty-fifth ratification) which requires the legal prohibition of national, racial, or religious hatred that incites discrimination, hostility and violence.

Definitions of hate speech vary across legal, scholarly, and platform contexts, creating inconsistent thresholds for recognising and addressing harm. While internationally recognised definitions emphasise identity-based hostility (United Nations, 2025), public and scholarly usage increasingly extends the term to encompass broader forms of offensiveness or hostility. Harms linked to hate speech include incitement to violence (United Nations, 2025), democratic polarisation (Terranova, 2022; Cachopo, 2022), psychological injury and diminished public participation among targeted individuals (Binny et al., 2020), although more work is needed to demonstrate exact harms both to individuals (Strossen, 2018) and to the utility of the digital environment as a place for reasonable political and social discourse.

Three of the most significant inconsistencies in how hate speech is defined, recognised and understood are as follows: Firstly, some definitions of hate speech limit it to content that disparages or harms people of the basis of racial, ethnic, or religious disparagement only. This narrow framing overlooks implicit hate speech such as sarcasm, coded language or context-dependent references (Ahn,

2024). In such cases, a statement may appear neutral in isolation but convey hateful meaning when interpreted within a broader conversational or social context. Many legislated definitions, including that by the European Union, do not include hate speech directed to or about ***gender, sexuality, gender identity, age or disability***, while these of course remain protected minorities in other jurisdictions (Carlson, 2021). Inconsistencies in what 'counts' as a minority or protected category of persons makes it difficult to navigate, report and remedy, or to articulate what kind of harm has been done.

Secondly, most (not all) laws that seek to prohibit hate speech, and some platform policy, is limited only to hate speech that has a high likelihood of ***incitement*** of physical violence, insurrection or serious unrest. While South Africa, Brazil and Germany have laws that may prohibit hate speech when it harms the dignity of protected categories of persons, other countries like the United States only act against hate speech when it has a very high likelihood of inciting imminent violence (Waldron, 2012). In the supranational space, the Rabbat Plan similarly limits the harm of hate speech to incitement (Carlson, 2021). The diversity in whether or not an incitement clause is used as a means to avoid curtailing freedom of expression, again creates substantial inconsistencies that stymie the fight to reduce the very high rates of hate speech.

Finally, research by some of the authors of this concept paper indicates that hate speech is widely adopted by everyday users to describe other kinds of digital harms or unwanted communication (insult, offensiveness and on some occasions even 'spam'). Our ethnographies of participants who have been subject to digital harms asked participants to code and submit the harms they experience or witness over a four-week period. Among those that were coded as "definitely hate speech" by participants, 36% were coded by the research team as other kinds of harmful or problematic communication, indicating a clear diversity of understanding among the community on the meaning and use of the term, and the fact that a post that is hate speech can also be something else.

At the same time, some scholarship has extended the term hate speech beyond discriminatory language or conduct in relation to established protected minorities (Waldron 2012), using the term to include offensive or hostile expression that falls short of established thresholds (Cinelli et al. 2021) and may include broad discrimination, humiliating, ridiculing, degrading or demeaning content or conduct (George, 2016; Mitts, 2025).

Together, this signals a shift in community expectations about what "counts" as harmful content or behaviour online (Cinelli et al. 2021). This widening usage blurs the boundary between hate speech and other forms of abuse, complicating moderation and policy responses. Platforms may remove explicit identity-based slurs but often struggle to detect coded, indirect, or euphemistic expressions, as well as harmful content that lacks clear identity markers (Walther 2022). The variation in legal definitions contributes to regulatory fragmentation, uneven enforcement, and difficulty for users knowing even if they have a case to report. The definitional atrophy also makes it difficult to manage or describe what a clean, non-toxic digital ecology might look like.

Cyberbullying

Cyberbullying is a term that emerged in the early 2000s to describe the practice of using digital communication channels to bully a person, such as by sending

messages or posting content of an intimidating, mocking, or threatening nature (Belsey, 2004). From early on, the term has tended to have a focus on harms to younger people, and often relied on an untested assumption that adults are resilient to online bullying while younger people are significantly more likely to be damaged by it (Fleming et al. 2006). The early scholarly use also tended to exclude the gender, sexuality and racial dimensions of abuse and harassment. Given the rise of social concerns about young people's safety online from the mid-2000s, the term was often incorporated into policy and educational materials, digital literacy and risk management curriculum, and in some early platform regulatory legislation. The Australian Online Safety Act 2021, for example, uses the term to describe harms to young people and differentiates it from harms experienced by those over the age of eighteen years, which is referred to instead as 'cyber-abuse'. Although more research is always needed on suicidality, cyberbullying has been connected with cases of suicide ideation and attempts among young people (Bauman et al., 2013; Hinduja and Patchin, 2010), and is obviously otherwise painful and hurtful to victim-survivors.

As with other digital harms terminology, cyberbullying is subject to diverse definitions. One study of the use of the term cyberbullying in legislation across European countries suggests very widespread definitional disagreement (European Commission, 2025) while another found that the lack of clarity just across Europe had prevented systematic research on the extent and impact of cyberbullying, making it impossible to draw conclusions on risk, protective factors and remedies at scale (Villar Onrubia et al., 2025).

The lack of clarity on definitions of cyberbullying is also made more problematic by continued assumptions about the nature and form of bullying itself. By framing online hostility and a broad range of digital harms as cyberbullying has limited community understandings about the effects and impacts of problematic digital content (Olweus and Limber, 2018). By invoking assumptions about older models of 'playground' bullying as well as very outdated ideas that such bullying is beneficial for the victims (McMahon 2005), it plays down the seriousness of digital harms.

When cyberbullying is used as a 'catch-all' for "offensive speech" and other harms (e.g., Hayes, 2017), it creates a highly confusing environment in relation to adults and younger people, particularly when it is assumed that only younger people are harmed online, or when it assumes that the perpetrators and victims of online harassment, hate and abuse are predominantly young people and that harms circulate among younger people. While some evidence of suicidality has rightly fostered a focus on young people and the resourcing of protective education, the term itself is outdated, the connection between adult experience and youth experiences needs further research, and the wider toxification of the digital environment through hostile, adversarial and hateful behaviours and targeted content affects all users, regardless of age. Our recommendation is that the term cyberbullying has so little currency today in regard to broad array of digital harms because it stymies the development of remedies and solutions pertaining to all users. This is not, however, to suggest that the term should be abandoned given its continued prominence in legal, educational and policy contexts; rather, that it should be treated as a legacy umbrella term that requires systematic mapping onto more precise categories of digital harm.

Trolling and rage bait

The term trolling began as a descriptor for early Internet behaviours that were considered playful and relatively harmless – catching out newbies in online groups through deliberate pranks and mischievous behaviour, such as pretending ignorance in order to gain an emotional response from newer users (Marwick and Lewis, 2016). In smaller group settings, trolling genuinely was good-humoured and not malicious. By the mid-2010s, however, the term was used to describe online behaviour and content designed to "disrupt and upset as many people as possible, using whatever linguistic or behavioural tools are available" (Phillips, 2015: 2-3). In the larger-scale setting of social media platforms, trolling became a malicious and adversarial behaviour, one which skirts under many platform policy guidelines and regulatory frameworks, but which genuinely causes upset.

A more serious subset of trolling that has become very common is rage bait. Rage bait is a term that emerged in informal online discourse to describe the manipulative practice of eliciting outrage and aggressive responsiveness in to increase profile traffic, subscribers, algorithmic ranking and in some cases online revenue. While the poster may or may not be expressing anger themselves about a topic, their content is deliberate and intentionally designed to invoke an outraged response or cause adversarial dialogue among other users (Cover, 2025). As a new manifestation of trolling, it is serious because there are health implications to encouraging other users to feel and experience rage on a regular basis.

Another variant has been identified as "gender-trolling," described as online violence and abuse that uses misogyny, (hetero)sexism, transphobia and anti-feminism to create digital settings as sites of online hate (Xie et al., 2022). This is an important consideration, given the extent to which misogyny and transphobia are core parts of experiencing a toxic online environment (Poland, 2016). But this also indicates the ways in which terminology changes and adapts with everyday usage, including when it is co-opted by scholars. Again, this creates divergent views – is trolling harmful pranking or the encouragement of damaging emotions or injurious speech used to spread hate and social division?

Trolling and its related terms tend not to be used in platform policy or regulatory frameworks, but they are important considerations within the umbrella of digital harms. Trolling and enragement may use hate speech or replicate forms of harassment to achieve their ends, indicating the common cross-over between digital harms that have typically been treated distinctively.

Online abuse and harassment

Online abuse and harassment are terms often used together to refer to threatening, humiliating, defamatory, coercive, or otherwise offensive behaviour on digital platforms (Haslop et al., 2021). Unlike hate speech, which is identity-focused, it encompasses a wide range of behaviours, including stalking, impersonation, image-based abuse (including deepfakes), doxxing, extortion, sexual harassment, rape threats, coordinated pile-ons and trolling (Bailey et al., 2021). Abuse involves the systematic misuse of power to control, coerce, or harm (World Health Organisation, 2002), while harassment entails repeated acts that create an intimidating or hostile environment (Einarsen et al., 2011). Crucially, it is behaviour-focused, applying regardless of relationship, intent, or whether harm arises from single incidents or sustained attacks, and it recognises the injurious effects on victims, bystanders and the wider communicative environment (Costello et al., 2019).

Definitions of online abuse and harassment differ across scholarship, law, and platforms, creating inconsistent thresholds for recognising harm. Key terms like abuse, harassment, cyberbullying, trolling, and hate speech are often used interchangeably, but with different emphases in law, policy and research. Abuse typically refers to harmful behaviour directed at another (World Health Organization, 2002), while harassment often implies repeated or persistent conduct (Bailey et al., 2021), although some scholars have argued that there is no agreed definition of harassment at all (Schoenebeck, Lampe and Triêu, 2023). Yet platform policies and legislation frequently blur these distinctions, leading to conceptual vagueness that can result in both under-enforcement (e.g., serious harms dismissed as “trolling”) and over-enforcement (e.g., satire or critique misclassified as abuse). Legal and platform definitions often rely on high thresholds such as “serious psychological harm” or threats of physical injury, excluding everyday insults, shaming, or persistent low-level hostility that nonetheless degrade users’ liveability online (see following sections). Feminist and critical scholarship highlights how these frameworks overlook cumulative and gendered harms, where repeated minor abuses can have significant psychological and social impacts (Jane, 2017).

One challenge of defining online abuse and harassment is that it is made more complex by the variety of targets. Minoritised groups are the subject of considerable online abuse and harassment when they are public figures as an attempt to silence them and remove their voices from public discourse (Phillips, Pathé and McEwan, 2023). Doxxing is a form of online harassment that occurs more readily in this context, as private details are unknown to abusers. Abuse and harassment can also occur within shared living spaces and care relationships, both between partners and between parents and children, roommates, or adult children and elderly parents (Levy and Schneider, 2020). This abuse is enabled by digital technology that supports a heretofore unimaginable level of access to information about the location, interactions and behaviour of

one's loved ones (McKay and Miller, 2021). Interpersonal stalking and harassment is poorly accounted for in typical cybersecurity solutions, which focus on outsider threats, rather than those close to (or inside the) home (Slupska, 2019).

As with other terminology, some scholars and advocates use the term harassment to cover a range of perpetrations that may affect targeted users. Cross (2019) considers online harassment to be coterminous with doxxing and swatting, while Geoghegan (2023: 1) refers to it as a "blanket term for online sexual harassment, cyberbullying, flaming, trolling, and cyberstalking." Again, while we can rely on courts to draw on decades of legal deliberation to make a determination in a case, when it comes to reporting, legislation and community awareness of what is acceptable and unacceptable content, behaviour, as well as whether an injury has or has not occurred, clear, agreed definitions are vital.

Volumetric abuse (Pile-ons)

Volumetric abuse, often referred to as a "pile-on," is an emergent term used to describe when large numbers of users direct criticism, shaming, hostility, or derision (whether coordinated or organic) toward a single target (Aghazadeh et al., 2019). Individually, posts within a pile-on may appear mild or innocuous, but in aggregation, they can produce severe harm, including psychological distress, reputational damage, and, in some cases, suicidality (Thompson & Cover 2022). Pile-ons can involve replies, quote-tweets, sharing, "liking" harmful content, or other forms of mass engagement intended to amplify social pressure (Aghazadeh et al., 2019). Unlike traditional content-based harms, volumetric abuse is intrinsically cumulative: its injurious effect arises not from the severity of each post, but from the scale, repetition, and simultaneity of mass participation. This makes it a distinct form of digital harm that reflects the networked, algorithmically-driven dynamics of digital communities.

Current policy and platform frameworks struggle to address volumetric abuse because they emphasise content-level assessment and evaluate posts *individually*, overlooking the aggregated impact of hundreds or thousands of hostile interactions (Nash & Felton 2024). As a result, volumetric abuse remains marginalised in taxonomies of digital harm, and pile-ons are routinely under-moderated despite victims experiencing them as profound injury, despite the significance of harm to individuals and its impact on public figures, many of whom have withdrawn from public life as a result of regular volumetric abuse (Cover, Henry et al., 2025). Volumetric abuse also blurs the line between legitimate critique and coordinated hostility, and current frameworks rarely account for the role of platform algorithms in amplifying and accelerating pile-ons, further intensifying their harmful impacts.

Disinformation, misinformation and malicious synthetic media

Disinformation, misinformation, and misleading content are central to contemporary digital harm frameworks. Disinformation, misinformation, malinformation and malicious synthetic media refer to false, misleading, or deceptively manipulated content that distorts information environments and undermines users' ability to make informed judgments. Disinformation refers to deliberately false or fabricated content created with the intention to deceive, manipulate opinion, or cause harm (Bennett & Livingston 2018). Misinformation, by contrast, describes false or inaccurate

content shared without intent to mislead, yet still capable of distorting public knowledge or producing harm (Wardle & Derakhshan 2017). Malinformation occupies an intermediate space: it may involve selective framing, distortion, or biased representation, and can at times overlap with defamation or reputational harm (Wardle & Derakhshan 2017). Malicious synthetic media encompasses AI-generated or AI-manipulated content (including deepfakes) designed to deceive, impersonate, or damage individuals, institutions or public trust (Meikle 2023; Cover 2022).

Definitions of misinformation, disinformation and malinformation frequently overlap, creating ambiguity for users, policymakers and platforms. The distinction between intentional deception (disinformation) and unintentional sharing (misinformation) is often difficult to establish in practice, particularly in fast-moving digital environments where users frequently share content without verifying accuracy. This makes regulatory or platform enforcement difficult, since the same piece of content can be classified differently depending on whether intent is assumed or ignored. Contextual factors further complicate definitional clarity. An accurate fact can become misleading when placed alongside unrelated or emotive material. Satire, parody, and memes blend humour with distortion, often evading categorisation altogether. Cross-cultural differences mean that content treated as misleading in one context may be seen as false (or even harmless) in another. Platform and regulatory definitions are similarly inconsistent and fragmented across jurisdictions, producing a lack of shared taxonomy and creating confusion for users and moderators about what qualifies as harm.

One critique of scholarly and research approaches to mis/disinformation is the way in which it is commonly silo-ed from other kinds of digital harms. While there is a logic to this, given the size and importance of mis/disinformation and its role in damaging electoral integrity, social harmony and health communication over the past decade, research is increasingly showing that it is deployed simultaneously and in support of other kinds of digital harms, including hate speech and abuse. Disinformation and misinformation frequently intersect with online abuse and harassment, operating as mutually reinforcing dynamics within digital ecologies. For example, climate change policy has been noted as significantly stymied by the double attack of aggressive hostility between advocates of different positions and substantial mis/disinformation on the topic (Kirkland 2025). False or misleading claims are often mobilised to legitimise hostility, stigmatise individuals or groups, and provoke pile-ons, while abusive practices amplify the visibility and circulation of disinformation through outrage, repetition, and algorithmic engagement. Research shows that harassment campaigns commonly rely on distorted narratives, conspiracy frames, or decontextualised content to justify targeting, silence dissent, and erode trust in authoritative knowledge. Conversely, abusive interactions create affective conditions – fear, anger, and polarisation – that make users more susceptible to misleading claims, enabling disinformation to spread more rapidly and persistently across platforms (Marwick and Lewis, 2017; Phillips, 2018; Wardle and Derakhshan, 2017).

Disinformation, misinformation and maliciously misleading content will always remain difficult to determine, because while some disinformation is obvious in many cases fact checkers, journalists and subjects of the content must clarify to

the veracity of the post, and then present that information in a way that mitigates the harm of misinformation (Hettiachchi et al., 2023). Nevertheless, having regard to clearer education for the wider population on the terminology is important, and has become even more so in an era marked by the mass-circulation of synthetic images and video (both deepfakes and AI-generated content), some of which has been used in government communication, including by the United States government (Harwell, 2026). Together, disinformation, unbranded synthetic image and video, hate speech, abuse, harassment and other harms play a role in supporting each in the toxification of the digital ecology and information environment, having a serious negative impact on the capacity of everyday users to utilise any content in exchange, entertainment, political deliberation and social engagement.

Digital addiction

Addiction has emerged in the wider discourse of digital harms to describe what is perceived or sometimes diagnosed as obsessive, continuous use of social media or gaming applications. In fact, an idea of internet addiction circulated long before social media, often incorrectly pathologising heavy use of the Internet and gaming (e.g., Young, 1998), mistakenly assuming that online engagement was an isolating rather than social activity (Grohol, 2000). Part of the problem with 'addiction' as a term is that it applies the rhetoric of drug addiction (pharmaceutical substance abuse) to the activity of digital media use, incorrectly assuming 'the digital' has similar effects as chemical dependence (Cover, 2004). It is also arguably a problematic term that applies an industrial-capitalist work ethic by pathologising procrastination and choices over the use of time in meaningless recreation (Thompson, 1967).

Nevertheless, there are realistic concerns in today's social media environment of the harm both to individuals and to the wider communication environment of platform feed algorithms, practices promoting adversarial discourse, non-finite short-form video scrolling feeds and disruptive notification frameworks that capture and maintain what Chris Hayes (2025) has called "involuntary attention." Attention is the key resource extracted by social media platforms and there is growing recognition that assumptions all attention is voluntary are incorrect. Involuntary attention is known to disrupt users' sleep routines, especially among young adults, and there is an argument that platform-designed mechanisms to retain attention are exploitative (eSafety Commissioner, 2024). Platform system designs and

algorithmic feeds can also prioritise other forms of digital harm (abuse, conspiracy theories, harassment and hate) due to their attention-grabbing nature and the ways in which they foster further engagement and sustained attention.

Capturing and retaining attention is a key aspect of most social media company's business models which, alongside harvesting data about users actively seeks to capture user attention and keep users on the platform for as long as possible (Wu, 2025; Carlson, 2021). Attention has long been recognised as a finite and scarce resource (Simon, 1971) and platform models have afforded the capture of attention in ways which are not always within the easy control of users themselves (Hindman, 2018). Artificial virality, speed of loading, and seamlessness between texts and posts, alongside the promotion of adversity and sensationalism as the principal form of promoted content on some social media platforms has enabled attention capture in its involuntary form (Kaye, 2019a).

Although involuntary attention and the wider operation of the attention economy are becoming better recognised among the corpus of digital harms, recent discourse calling for governments, platforms and communities to address the issue has unfortunately returned to the older discourse of digital addiction. For example, Jonathan Haidt's (2024) book *The Anxious Generation*, which was widely cited as a justification for Australia's so-called social media age ban despite flawed evidence, refers regularly to young people's 'addiction' to social media and mobile devices. The Australian Government's Department of Communications and the Arts refers to the problem of 'addiction' (Australian Government, 2025). In 2024, New York State passed a law 2024 to reduce children's addiction to social media, while California passed the *Protecting Our Kids from Social Media Addiction Act* which will come into effect in January 2027. British politicians have also raised the need to address 'addiction' (Stokel-Walker, 2024).

There are a number of problems that emerge with the definitional choice to refer to involuntary attention as 'addiction'. Following public discourse about other kinds of addiction, including substance abuse, there is a tendency to assume excessive time on social media is a problem over which individual are solely responsible and over which they have agency (LaRose, Kim and Peng, 2011), thereby ignoring the causal effects of algorithmic feeds deliberately designed to attract and sustain attention. The addiction discourse also assumes that remedies will be individualised, drawing attention away from the important work necessary to address design issues, and the impact they are having on other aspects of social engagement. We believe a discourse of *involuntary attention capture and extraction* is far more useful, not only to avoid the problematic individualisation and pathologisation that often accompanies addiction and dependency discourses (Cover, 2004), but as a mechanism to recognise the damage systems designed to profit directly from attention may be causing both to individual users and the wider digital ecology.

Offensive content

Another set of terms found across some legislation, much scholarship and in community advocacy for a safer Internet is "offensiveness" and "offensive content." Much scholarship, for example, uses the term "offensive content" and advocates for better regulation to restrict its circulation without explaining or delimiting the term.

There are clear distinctions, however, that need to be foregrounded. For example, offending another user – through insulting them or ridiculing them (Eribon, 2004) – is distinct from posting something that is 'offensive' to liberal-minded audiences – such as saying 'Hitler had some good ideas'. One should be restricted because it may cause injury, while the latter is an expression that in most contexts is ignorant. However, there are contexts in which the latter phrase may be deeply offensive: if, for example, posted directly or tagging or otherwise deliberately drawing the attention of a holocaust survivor.

The grey areas of offense may thus also be dependent on contexts outside any audience (whether targeted or encountering recklessly). For example, circulating footage of a mass shooting may be offensive when done during a period of public grieving, while the same footage circulated online to warn others, as news, or in order to engage in analysis may not meet any threshold of offensiveness. As some platforms, such as Discord, permit "offensive content" to be reported (Cover, Beckett et al., 2025), it is unclear the extent to which either automated or human-managed moderation is 'trained' to unpack the contextual distinctions in what counts as offense.

One of the reasons why its use is problematic is because there are various supranational expectations that offensiveness is never in itself to be banned. Kaye (2019a), for example, notes that while Article 20 of the International Covenant on Civil and Political Rights (1966) seeks to restrict advocating hatred or incitement to violence, Article 19 protects freedom of expression. This is to be interpreted as meaning that as long as person is not advocating hate, discrimination or violence, freedom of expression warrants that offensive material is should not be restricted. Kaye uses the example of the offensiveness of a minority position, such as the interpretation of a religious tenet.

Claims to offensiveness have been used in many ways not only to restrict freedom of expression but to punish minorities – for example, claims that transgender people's existence is offensive, or the generation (dog-whistling) of mass indignation at a reasonable statement or creator by pretending offense – that is, playing the provoked victim (George, 2016).

Most legislation that incorporates offensiveness refers to the 'reasonable person' test as a way of differentiating between what offends an individual and what might offend the community. For example, words of blasphemy such as "my god" may offend some people but such words are not restricted by law or policy when they are regularly used across communities.

Offensiveness, then, becomes another area where leadership in providing definitional intelligibility is necessary to generate clarity in use, and to avoid the concept's misuse to restrict free speech. Some scholars have argued that a clearer distinction between hate speech, deliberate abuse and 'offensive content' is necessary, whereby that which offends but does not injure the dignity of a person or group of people may be *upsetting* and even immoral but should not be restricted (Carlson, 2021; George, 2016; Waldron, 2012). On the other hand, causing upset to users repeatedly may have emotional and psychological impacts that, over time, genuinely injure. In this respect, there is a simultaneous need to consider what kinds of offensiveness should circulate as freedom of expression and what kinds may genuinely cause harm.

3.2 Legal and jurisdictional differences

Legislation plays an important role in defining key terminology and is expected to guide not only regulatory practices, administrative decision-making, judicial judgments and lawyers, but to provide a reference point for members of the community and to help shape community standards through clarity. Much of the pressure to regulate and improve safety across digital platforms has come from governments through legislating regulatory measures, administrative requirements for transparency reports on content moderation, and requirements to remove specific pieces of highly problematic content such as disinformation, depictions of violence acts and significant hate speech (Mitts, 2025). In many cases, legislation in different jurisdictions such as Australia, the United Kingdom, Germany, the European Union, some states in the USA and other settings have provided definitions for different digital harms. These are usually brief and minimalist with the assumption that courts will interpret and determine meaning in application if and when required.

The fact that these regulatory initiatives were legislated in different counties and states at different times has resulted in very wide differences in legal across jurisdictions, producing a fragmented and often inconsistent regulatory landscape (Vincent 2017; Nash & Felton 2024). Governments use different terms – such as abuse, harassment, hate speech, or cyberbullying – and apply different thresholds for what counts as harm. Some laws require proof of "serious physical or psychological harm" for a content or behaviour to be considered harmful, while others recognise broader impacts such as "emotional distress" or sustained "campaigns of mistreatment". Most frameworks remain focused on individual pieces of content and whether they cause harm within a particular jurisdiction, rather than having the capacity to recognise cumulative harms, patterned behaviours, or damage to the broader digital environment (Vincent 2017). Jurisidictional distinctions between perpetrator known wrongdoing (*mens rea*), recklessness and negligence also further complexify the legal setting for 'making sense' of digital harms internationally.

These differences make it hard for people to understand their rights and protections online, and equally difficult for platforms to apply consistent standards across countries. This fragmentation reflects what Flew (2021) describes as the re-nationalisation of internet governance: national laws attempting to regulate a global, post-national communication environment. As a result, the same behaviour may be lawful in one jurisdiction, unlawful in another, and inconsistently moderated on global platforms. A further source of unintelligibility arises from the mismatch between statutory definitions of harm and the ways platforms define and enforce their own rules. Legal frameworks tend to emphasise demonstrable individual injury, while platform policies focus narrowly on content or discrete behaviours. Neither approach adequately accounts for cumulative toxicity or harms to the broader digital ecology. To illustrate this fragmentation, Table 1.0 compares how three common-law, English-speaking jurisdictions define and regulate harmful digital communication in legislation.

Table 1.0: Comparison of digital harm legislation across three common-law jurisdictions.

Jurisdiction	Key legal framework(s)	Definition / threshold of harm	Distinctive features	Challenges / implications
Australia	*Online Safety Act 2021*	Cyberbullying (minors): material that is "seriously threatening, seriously intimidating, seriously harassing or seriously humiliating." Cyber-abuse (adults): must be likely to cause "serious physical harm or serious harm to mental health" and be material that "an ordinary reasonable person would regard as menacing, harassing or offensive in all the circumstances."	High threshold; community-based reasonableness standard; age-based distinctions.	Excludes dignity harms and everyday toxicity; limited recognition of cumulative or environmental harms; hard to prove intent in digital settings.
United Kingdom	*Online Safety Act 2023*	Bullying content in relation to children includes material that "conveys a serious threat," is "humiliating or degrading," or forms part of a "campaign of mistreatment."	Broader recognition of psychological, reputational, and campaign-based harms; acknowledges sustained mistreatment, not just single posts.	Still largely content-focused; limited treatment of algorithmic or systemic amplification; context and cumulative effects not fully integrated.
New Zealand	*Harmful Digital Communications Act 2015*	Harm is defined as "serious emotional distress" and requires demonstrating that harm actually occurred.	Lower threshold than Australia or the UK; contextual and victim-sensitive assessment; recognises varied resilience and vulnerability.	Burden placed on victims to prove distress; enforcement depends heavily on individual context; difficult to apply consistently across diverse user experiences.

We have noted this small sample of legislation from three countries to point to the similarities and significant divergences of how digital harms terminology is defined, noting that the divergences are likely to be exponential across a much wider jurisdictional pool. Notably, each uses different terms (cyberbullying, bullying, harassment), and in each case provides notably different thresholds for harm and different means by which to determine harm. Here is where the re-nationalisation of the Internet (Flew, 2021) exacerbates unintelligibility in a digital setting founded on post-nationalist, global interaction.

3.3 Platform definitions of digital harm

Operating as private companies, platforms govern user behaviour and content through two sets of rules: terms of service and community guidelines, together often referred to as "platform policy" (Jiang et al., 2021). Despite having their own policy frameworks, platforms must also comply with the range of laws discussed above, often on the regulatory expectation that they will police, censor, intervene and take down content on behalf of the states where it has occurred, or where a user is residing (Kaye, 2019b). As social media platforms are considered private virtual places in a contractual relationship with their users or customers, they are not bound by the United States' First Amendment requirements to protect freedom of expression, and thereby have greater latitude to apply rules to the content and created or shared on their services (Carlson, 2021).

As with jurisdictional and scholarly definitions, digital platforms also use highly variable and often opaque definitions of online harms, *creating another layer of inconsistency for users and regulators*. Most platforms define harmful behaviour through the lens of prohibited *content* or *conduct* – such as bullying, harassment, hate speech or threats – but rarely define *harm* itself and are broadly disinterested in determining harms by injury, only by a standard of content or reported behaviour (Cover, Simcock and Humphries, 2025).

Instead, platforms rely on narrow, content-focused criteria that emphasise individual posts, slurs, or explicit threats, while overlooking cumulative injury, contextual meaning, or the experiences of victims and bystanders. Some platforms frame harms primarily as behaviour (e.g., "targeting," "inciting harassment"), whereas others use broad and sometimes vague descriptors such as "intimidation," "distress," or "shutting someone out of a conversation." Definitions are frequently scattered across multiple policy pages, inconsistently applied, or updated without clarity, making them difficult for users (and even moderators) to navigate (Cover, Beckett et al., 2025). Research has shown that platform application of policies are deliberately and knowingly inconsistent (Flynn, Vakhitova et al., 2025; Cover, Henry et al., 2025)

Crucially, platform frameworks typically ignore the role of algorithms in amplifying toxicity, and rarely consider the wider digital ecology or collective harms such as pile-ons or coordinated campaigns (Orton-Johnson, 2024). As a result, platform definitions do not align with one another, with legal frameworks, or with everyday user experiences of harm, creating confusion and limiting the development of a coherent, shared taxonomy of digital harms. Table 2.0 provides a comparative snapshot of how major platforms articulate and operationalise bullying and harassment.

Table 2.0: Comparison of digital harm policy across major social media platforms

Platform	How harm is defined / key policy language	Content-focused or behaviour-focused?	Distinctive features	Challenges / implications
Meta (Facebook / Instagram)	Tiered "bullying and harassment" policy prohibiting content intended to "degrade or shame," as well as repeated unwanted contact, sexual harassment, or behaviour directed at large numbers of individuals without prior solicitation (Meta, 2025).	Mostly content-focused, with some behaviour recognition.	Tier system creates hierarchical categories of harm; stricter rules for minors.	Emphasis on individual posts means cumulative or contextual harms (e.g., pile-ons) remain largely invisible.
YouTube	"Harassment & cyberbullying" policy prohibits "prolonged insults or slurs" based on intrinsic attributes, as well as threats, doxxing, and (for minors) content intended to shame, deceive, or insult" (YouTube, 2025).	Content-focused.	Stronger protections for minors; focuses on slurs and offensive content.	Narrow content categories overlook patterned behaviour or systemic amplification.
Reddit	Defines harassment, threats, or bullying as "anything that works to shut someone out of a conversation through intimidation or abuse" (Reddit, 2025).	Content + behaviour.	Unique focus on conversational exclusion and intimidation.	Broad, undefined terms ("anything," "shut out") result in uneven moderation and moderator discretion.
X (formerly Twitter)	Prohibits "targeted harassment," "incitement of harassment," unwanted sexual content and "graphic objectification," and "insults" (X, 2024).	Primarily behaviour-focused.	Recognises linking/tagging behaviours as harassment; acknowledges malicious targeting.	Minimal content restrictions; high tolerance for harmful but lawful speech allows for significant toxicity.
TikTok	Prohibits "harassing," "bullying," or "degrading" behaviour (TikTok, 2025a); defines bullying as "targeted behaviour" that intends "physical, social and/or psychological "harm" (TikTok, 2025b).	Behaviour-focused.	Some recognition of social and psychological injury.	Behaviour is still inferred through content-based evaluation.

An alternative way to understand how platforms define, conceptualise, determine and act upon digital harms is differentiated by ***content-focused models*** versus ***behaviour-focused models***.

Content-focused approaches, exemplified by Meta, YouTube, and Reddit, prioritise the assessment of individual posts, images, or messages against predefined categories of prohibited material. Meta's tiered schema evaluates the harmful qualities of speech itself, while YouTube prohibits specific types of insulting, shaming, or threatening content, particularly where minors are concerned. Reddit similarly frames harassment as "anything" that intimidates or excludes, implicitly tying harm to textual or visual material. Across these platforms, harm is primarily apprehended as something *contained within content*. The text argues that this model is increasingly inadequate: perpetrators can exploit moderation thresholds by producing superficially mild content that nonetheless functions strategically to provoke or undermine others. Content-based regulation also struggles with legal constraints, particularly in the United States, where free-speech protections and intermediary liability regimes shield platforms when harms are defined narrowly as speech. It is also limited by the practice of content moderation where very quick decisions are assessed based on an individual employee's prior knowledge or an algorithm's prior programming (Gillespie, 2018). Content is usually treated as individual, one-by-one pieces and assessed or policed outside a behavioural context and without reference to testimony of an injured party, thereby constraining the definition of digital harms to the text itself.

By contrast, behaviour-focused approaches, evident in X, TikTok, Discord, and LinkedIn's policies, conceptualise harm as actions directed at individuals. These policies emphasise targeting, repetition, harassment, and incitement, even where offensive content is tolerated. Here, harm arises not from what is said alone, but from how content is used – through tagging, piling-on, or sustained intimidation. However, this approach individualises harm, struggles to assess intent or repetition, and is ill-suited to misinformation, where damage often stems from circulation rather than deliberate perpetration.

Overall, the divergence between content and behaviour frameworks produces regulatory confusion and user uncertainty. In addition to the scholarly and legislative divergences in definition, this underscores

the need for more coherent, systems-based understandings of digital harm.

3.4 Emergent definitional alternatives

While there have been some limited attempts to develop cross-sectoral taxonomies described in 2.2 above, there are is a clear need urgent need and a legitimate argument for definitional clarity and systematisation of the meaning of the full range of terminology associated with digital harms and online safety in cross-sectoral, interjurisdictional and interdisciplinary contexts if a discourse towards systems-based remedies and duties of care are to be invoked and adequately co-created with users, victims and stakeholders.

Some alternative ways in which to consider, understand and apprehend digital harms have been invoked in recent scholarship. Jiang and colleagues (2021), for example, have noted that ***severity of impact*** may be a suitable mechanism to determine digital harms, eschewing the distinctions between, say, disinformation, abuse, trolling, bullying and offense. A taxonomy based on severity of impact, which may include injury or likelihood of injury, could, according to the authors, guide content moderation and automated processes through prioritising interventions. It may also help platforms, jurisdictions and researchers to determine and predict where 'severe' digital harms are more likely to occur and thereby help design better preventative and systems-based tools.

Others (e.g., Thomas et al., 2022) have suggested ***virality*** may be a better determinant of a harm or cluster of harms than any strict definition given in platform policy or legislation. For example, a hateful comment that is encountered within a certain period by only a small number of users may be considered less harmful than a hate comment that circulates to thousands or millions of users. Again, considering the impact of virality may provide better frameworks for how a digital duty of care is described to platforms and/or serves as a regulatory lever to create a less-harmful digital setting – for example, platforms could be required to undertake a duty of care not by censoring content that may be ambiguous, and avoiding shadow-banning as a secretive technique, but by deliberately reducing the virality of particular content as a 'risk' that it may fall into a particular kind of harm.

> "severity, virality, intent and injury may contribute to a matrix of digital harms taxonomy that helps shift beyond the existing impasse"

Some scholars have offered alternative definitions and conceptualisations based not on definitions that focus on content or how it circulates, but on ***perpetrator intent***. Intent is always difficult to police, but there is value in considering how a hierarchy of intent may help to define what is a digital harm. Walther (2025), for example, notes the significance of intentional versus reckless violence, and this leads to possibilities to consider how a hierarchy of intentional, reckless and negligent harms may be deployed as ways

of determining if content or behaviour is harmful, regardless of third-party assessment of the content or behaviour itself.

difficult to demonstrate, including in the context of legal action. However, a more precise yet expansive definition of what constitutes an injury may turn out to be a better way of categorising, taxonomising and framing digital harms than extant definitions.

> "Intent is always difficult to police, but there is value in considering how a hierarchy of intent may help to define what is a digital harm"

Finally, a number of other approaches have argued that demonstrated ***injury*** may be the most useful approach not only to determining if content or behaviour is harmful, but to defining digital harms (e.g., Schoenebeck, Lampe and Triêu, 2023; Cheng et al., 2017; Heng et al., 2017). Waldron (2012), for examples, argues that we can overcome the enormous difficulties of defining hate speech, including in terms of differences on who is part of a protected category or community, by understanding it as an injury to dignity in much the way inequitable or unfair discrimination is defined. Much legislation seeks – variably – to define digital harms in terms of demonstrated injury, such as psychological or emotional injury (Strossen, 2018), which of course can be

While severity, virality, intent and injury are all known in case law to be difficult to determine exactly, they may contribute to a matrix of digital harms taxonomy that helps shift beyond the existing impasse caused by substantial definitional uncertainty across the domains of scholarship, statutory law and platform policy, and the resulting confusion for the wider community of users, including potential perpetrators.

4.0 Refiguring harms: Individuals and the digital ecology

Understanding and responding to digital harms requires a level of conceptual clarity that is currently missing from public, scholarly, legislative and platform discourse. To support this, any taxonomy of digital harms must operate across two interconnected domains. The first is the ***interpersonal domain***, where individuals experience harm through hate, trolling, abuse, misleading information, misrepresentation, shaming or humiliation (Cefai 2020) – harms that can lead to social withdrawal (Poland 2016), anxiety, stress or depression (Hinduja & Patchin 2014), family or professional impacts (Mosco 2017) and suicidality (La Sala et al. 2024). This is where there is a responsible person or people who are perpetrators to be held accountable, a responsible person or people who are victims or victim-survivors of those harms, and responsible person or organisation that facilitates, communicates, shares or amplifies the harm.

The second is the domain of the ***digital ecology***: the broader communicative environment that sustains social connection, work, political participation and cultural life. Drawing on ecological thinking as a system in which interdependent actors shape one another's conditions for liveability (Haraway 1990; Berlant 2022), we emphasise that digital harms also injure the communicative setting itself – through algorithmic amplification, platform design choices, and cultural practices that produce toxicity, distortion or exclusion, in ways that **actively toxify the wider digital setting** (and its cultural meanings) as a place for deliberation, social engagement, pleasure, work, creativity information. It may do this through the permanency of hate content or the virality of disinformation, or through the fear generated by knowledge of that toxicity that stymies creative engagement online and reduces mutual care not among individuals but a wider public. Together, these domains highlight that digital harms affect both people and the environments in which digital life operates.

"Understanding and responding to digital harms requires a level of conceptual clarity that is currently missing from public, scholarly, legislative and platform discourse"

Most existing approaches to digital harm concentrate on the immediate impact on individuals. That focus remains essential, but it does not fully reflect how today's digital environments operate. Digital platforms function as shared environments in which millions of people work, socialise and express themselves (Gillespie, 2018). When harmful content or behaviour circulates within them, it affects not only individuals but also the wider setting in which digital life takes place, reflecting how violence damages the shared conditions that support collective life (Butler, 2020). A clearer understanding of digital harm therefore needs to account

for both levels: the impact on individuals and the health of the digital ecology as a whole.

4.1 The individual

Individuals experience digital harm differentially. Social position, resilience, access to support and online visibility all shape how severely harmful content or behaviours affect them (Unger 2012). Abuse, harassment and large-scale pile-ons can cause deep psychological and reputational injury, including anxiety, depression, withdrawal from online participation and, in extreme cases, suicidality (Thompson & Cover 2021). These harms are well documented (Bauman et al. 2012; Hinduja & Patchin 2010; Young et al. 2017) but remain inconsistently defined across platform policies and regulatory frameworks, making them difficult to remedy.

Importantly, not all harmful experiences fit neatly into existing legal or platform categories such as hate speech or abuse. Many users face emotional harms that sit below formal thresholds (Harvey, 2017) but still produce significant consequences, including anger, shame, humiliation and withdrawal (Cefai, 2020; Keighley, 2022). These harms disproportionately affect women, gender minorities and racial and ethnic communities (Salter & Blodgett, 2017; McRobbie, 2020). A robust taxonomy of digital harm – and any concomitant duty of care requirement placed on platforms – should capture this broader spectrum, recognising not only deliberate attempts to injure but also behaviours and systems that disrupt a person's sense of safety, dignity and liveability online.

Harmful experiences are also not defined simply by one-to-one interactions. Rather, both incidental (e.g. social media 'pile-ons') and intentional (coordinated targeting of an individual) actions (Mariconti et al, 2019) can seem on a one-by-one basis to be insubstantial, but in aggregate lead to significant impacts on the person on the receiving end (Keighley, 2022). These circumstances are a challenge to models that only look at the individual-to-individual context. Other harms emerge from a mixture of online and in-person relationships, as in the context of partner coercive control (McKay and Miller, 2021). Interpersonal connections can add a dimension to digital harms that an exclusively online focus often overlooks.

4.2 The digital ecology

Digital harm is not limited to the direct experiences of individuals. The environments in which people interact – the digital ecology – can themselves become degraded in ways that harm collective wellbeing (Koo et al., 2024). When hostility, manipulation, misinformation or exploitative design practices become widespread, they make digital spaces less usable, less trustworthy and less supportive of communication, creativity and work. This structural damage is increasingly visible across social media, gaming communities, comment sections, and other online platforms, where hostility and disinhibition are well documented (Wachs & Wright, 2018; Keighley, 2022). Such conditions silence or marginalise the voices of already vulnerable groups, undermining the inclusivity and democratic quality of digital public life.

An ecology is usually defined as the supportive setting for organisms to interact with their environment and with one another, including the geographic, infrastructural, human and nonhuman collectivities that provide the conditions for sustainability and cultural evolution

> “digital harms also injure the communicative setting itself – through algorithmic amplification, platform design choices, and cultural practices that produce toxicity, distortion or exclusion, in ways that actively toxify the wider digital setting”

(Mead, 2017). Ecologies entail subjects that are interdependent with one another and vulnerable to one another as an irreversible condition of liveability, calling for an ethical responsibility to sustain that ecology for the cohabitation of all. A digital ecology also includes not only the platform that enables sharing or the people who engage there interactively but knowledge frameworks, cultural meanings, infrastructural supports, digital workers, engineers, protocols, technologies, and the wider physical and natural environment that may be impacted by this structure (Berlant, 2022). Knowledge frameworks that enable thinking about the protection of the physical ecology are long-standing and well-recognised, and capable of determining harm to an environment separate from injuries to individual human beings. In health, socio-ecological models have also been a long-standing framework for understanding human behaviour (Bronfenbrenner, 1979). They posit that individual health behaviour is a product of how individuals interact with multiple actors in the ecology, which can have both intentional and unintentional influences on behaviour. These models have been applied to the digital context, to understand how interactions with digital platforms can impact health behaviour (Micallef et al. 2022). A similar framework, we argue, can apply to overcome the lack of definitional clarity and conceptual impasses we have described above.

A taxonomy that recognises harm to the digital ecology broadens the focus from individual complaints to the health of

the system as a whole and helps move online safety away from questions of vulnerability and risk to questions of the toxification of the digital ecology. It recognises that bystanders are affected even when they are not the direct targets of abuse or disinformation: they encounter hostile exchanges, see harmful narratives amplified, or adapt their behaviour to avoid becoming targets themselves. It also recognises that future users, who may not yet be born, will be affected by present rates of toxicity in online settings through the permanence of disinformation and the persistence of hate and abuse cycle. It recognises that how we see ourselves, choose food, entertain one another and engage in play – all essential activities for liveability – are conditioned by a wide ecology that is sometimes difficult to delimit (Micallef et al., 2024) And it recognises that a toxic environment over time normalises incivility and adversity as a way of behaving and engaging, reduces participation, and erodes the shared norms that make digital environments functional. Accounting for ecological harm allows policymakers, platforms and communities to think beyond reactive enforcement and towards sustaining environments where social, civic and economic digital life can thrive.

A focus on the harms to the digital ecology also allows us to avoid weighing up the differences, intents, contexts and range of injuries of individual perpetrations in an interpersonal context. For example, whether hate speech is likely to incite violence, whether or not disinformation was intended, if rage baiting has been produced by an automated process, and if pile-ons should be treated as individual mild content or as volumes of mass shaming, whether orchestrated or not. That is, looking to the harms to the digital ecology as a necessary, vital space for communication, information, social interaction, global agonism, political debate, creativity and economic activity (Wu, 2025) enables us to define digital harms in terms of violences towards the social order (Engels, 2015) that depends today so centrally on digital platforms.

5.0 Directions for policy, research and practice

Building a coherent response to digital harms requires coordination. The definitional fragmentation identified in this paper highlights the need for shared language, interdisciplinary methods, and governance approaches that recognise harms to both individuals and the wider digital ecology. The following directions outline priority areas for future policy, research and practice.

tools are needed – drawing on human-computer interaction, computer science, public health, psychology, humanities, social sciences, philosophy, business, and legal studies – to capture cumulative harms, ecological harms, algorithmic amplification, and the contextual nature of injury in digital environments.

The purpose of the proposed taxonomy is not to optimise content moderation, but to support prevention, safety-by-design, and platform duties of care by clarifying what kinds of harm are occurring and how they accumulate within the digital ecology. While dimensions such as severity,

> "Approaches to definition and to related duties of care, must therefore acknowledge that online harms injure both people and the communicative environments they inhabit"

5.1 Develop a multi-sector, agreed taxonomy of digital harms

A first-order priority is the creation of a shared, multi-sector taxonomy that can be used consistently across platforms, jurisdictions, regulators, researchers, communities, employers, journalists and fact-checkers. A common vocabulary would enable more intelligible reporting, clearer regulatory obligations, and greater consistency in moderation and remedy, including in the interjurisdictional setting.

5.2 Build interdisciplinary tools and methods for measuring harm

Current measurement practices rely on narrow indicators such as explicit threats or discrete abusive posts. More robust

scale, and virality may inform moderation decisions as a downstream application, the primary function of the taxonomy is upstream: to guide design choices, risk assessment, accountability frameworks, and preventative interventions, rather than to refine post hoc content removal.

5.3 Ground definitions of digital harms in ecological harm first, and individual harm second

In this framework, ecological harm provides the primary organising principle, while dimensions such as intent, severity, and scale function as secondary descriptors used to guide prioritisation, proportional response, and duty of care, rather than to determine whether harm exists. This prioritisation is not

“Regulatory and platform frameworks must evolve to match the ways digital life has changed”

intended to diminish the importance of interpersonal or individual harm, but to address a longstanding tendency in policy and platform governance to treat harm primarily as isolated incidents having a measurable impact on individuals, rather than as cumulative, systemic, and ecological phenomena.

Definitions of harm must expand beyond narrow thresholds to reflect contemporary realities, and recognise that the future of the digital setting is closely entangled with the future of humanity, of politics, of society and of the physical world. This includes recognising harms to the digital ecology and to human dignity broadly, rather than relying on liberal-humanist assumptions that prioritise the individual. Such a position is not to do away with definitions that focus on individuals within communicative chains, but to recognise that harms to dignity, safety, trust, privacy and reputation are also always harms to the digital ecology in terms of structure, infrastructure, civic cohesion and interdependency as an a priori condition of survival. Approaches to definition and to related duties of care, must therefore acknowledge that online harms injure both people and the communicative environments they inhabit.

5.4 Develop twenty-first century understandings of harm

Regulatory and platform frameworks must evolve to match the ways digital life has changed. Offence, humiliation, reputational injury, civic degradation, and the toxification of digital spaces may be more harmful today than in pre-digital contexts. New conceptual frameworks are needed to understand and address these forms of injury, including the ways distrust, polarisation and systemic toxicity function as harms in and of themselves.

5.5 Introduce or strengthen platform duties of care

A shared, international duty-of-care framework could provide a unifying governance mechanism. Such a framework would require platforms to anticipate, mitigate, and report risks – including cumulative harms, algorithmic amplification, and damage to the digital ecology – rather than focusing solely on removing individual pieces of content. A coordinated duty of care would also help align platform practices with regulatory expectations and user needs across jurisdictions.

6.0 Conclusion: Building a healthier digital ecology

This paper has argued that digital harm must be understood in dual terms: as the injury experienced by individuals and as the wider degradation of the digital ecology that sustains social, cultural and political life. Current approaches remain limited by inconsistent definitions, narrow content-focused frameworks and a tendency to treat harm as an isolated interpersonal event rather than a systemic and ecological phenomenon. A shared taxonomy of digital harms offers a way to bridge these gaps. It would enable clearer communication across sectors, support more consistent regulatory design, and allow platforms to recognise and address harms that accumulate, amplify and spill across environments.

“Current approaches remain limited by inconsistent definitions, narrow content-focused frameworks and a tendency to treat harm as an isolated interpersonal event rather than a systemic and ecological phenomenon”

We have focused on those harms most ostensibly related to violence and injury of parties, to harms to the digital ecology, and for which we see the greatest need for clarity to enable regulatory clarity,

legal efficiency, genuine duties of care, the production of mutual care among users, and protecting and detoxifying the broad digital ecology. There are dozens of other online harms that also need attention: scams and fraud, including those which rely on synthetic media, AI bias, the use of spam bots to flood feeds or to manipulate public opinion about international relations, digital vandalism of sites and posts, broad privacy issues and the security of private information held by corporate actors and governments, and so on – the list grows longer. We argue that while there is sometimes a scholarly or regulatory need to address each individual 'type' of potential harm individually, there is enormous benefit in building a taxonomy under the umbrella of "digital harms" in a way that draws attention to the toxicity of the digital environment and the concomitant individual and social damage that engaging in such an environment can do.

Developing such a taxonomy requires deliberate interdisciplinary collaboration, integrating insights from human–computer interaction, computer science, health and medicine, social sciences, philosophy, business and legal studies. It also demands attention to both measurable injury and the conditions of the digital ecology itself – the algorithms, cultures and infrastructures that shape collective life. Future work should focus on refining core definitions, establishing common measurement tools, and exploring how an international platform duty of care could embed ecological responsibility into governance. By pursuing these directions, policymakers, researchers and platforms can contribute to a healthier, more resilient digital environment in which individuals and communities are able to participate safely and meaningfully.

"Developing such a taxonomy requires deliberate interdisciplinary collaboration, integrating insights from human–computer interaction, computer science, health and medicine, social sciences, philosophy, business and legal studies"

7.0 Recommendations

There are four key recommendations for governments and legislators, scholars, platform stakeholders, community advocates.

Recommendation #1:

Governments, platforms, scholars, health practitioners, lawyers, educators and community advocates work together globally, led by supranational organisations, to develop a clear, shared, agreed taxonomy of digital harms within five years. The taxonomy will be flexible and adaptable to account for emergent harmful content, behaviours and practices.

Recommendation #2:

That the taxonomy (a) be determined by injury and harm, (b) recognise the intersection of different harm 'types', and (c) prioritise harms and injuries to the digital ecology rather than individual and interpersonal harms.

Recommendation #3:

That a shared, agreed and internationally-relevant taxonomy underscore and drive the development of safety-by-design, digital duties of care, and education for children and adults that emphasises why harmful practices, posts, content and behaviours may be injurious to others, the digital ecology, the community and the future.

Recommendation #4:

Develop better understandings and regulation of the inter-platform environment and the interjurisdictional post-national Internet, and requirements on platforms and governments to work together to address and prevent harms that occur across platforms, and across jurisdictional boundaries. Ensure new definitions of digital harms recognise the incompatibility with jurisdictional and platform definitions, and build safety-by-design principles and duties of care that recognise communication takes place beyond and across jurisdictional boundaries and corporate platform policy and intervention techniques.

8.0 References

Aghazadeh, S. A., Burns, S., Chu, J., Feigenblatt, H., Laribee, E., Maynard, L., Meyers, A. L. M., O'Brien, J. L., & Rufus, L. (2019). GamerGate: A case study in online harassment. In J. Golbeck (Ed.), Online harassment (pp. 179–207). Springer.

Agrafiotis, I., Nurse, J. R. C., Goldsmith, M., Creese, S., & Upton, D. (2018). A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate. Journal of Cybersecurity, 4(1), 1–15. https://doi.org/10.1093/cybsec/tyy006

Aleti, T., Figueiredo, B., Reid, M., Martin, D.M., Sheahan, J., Hjorth, L. (2025). Older adults' digital competency, digital risk perceptions and frequency of everyday digital engagement. Information Technology & People 38(8), 97-118. https://doi.org/10.1108/ITP-05-2024-0624

Amnesty International (2018b). Online abuse of women thrives as Twitter fails to respect women's rights. Amnesty International, 20 March. https://www.amnestyusa.org/reports/online-abuse-of-women-thrives-as-twitter-fails-to-respect-womens-rights/

Ahn, H., Kim, Y., Kim, J., & Han, Y.S. (2024). SharedCon: Implicit Hate Speech Detection using Shared Semantics. Findings of the Association for Computational Linguistics, 10444–10455. Association for Computational Linguistics. https://aclanthology.org/2024.findings-acl.622.pdf

Australian Government (2025). Social Media Minimum Age – Fact Sheet, July. https://www.infrastructure.gov.au/sites/default/files/documents/social-media-minimum-age-and-age-assurance-trial-fact-sheet-july-2025.pdf

Bahl, R., McKay, D., Chang, S., Buchanan, G., & Cheong, M. (2025). The phantom information booth: migrant and sedentary tertiary students' tactics in the face of suspect information on social media. Journal of Documentation,81(2), 526-544. https://doi.org/10.1108/JD-06-2024-0153

Bailey, J., Flynn, A., & Henry, N. (Eds.). (2021). The Emerald international handbook of technology-facilitated violence and abuse. Emerald.

Bauman, S., Toomey, R. B., & Walker, J. L. (2012). Associations among bullying, cyberbullying, and suicide in high school students. Journal of Adolescence, 36(2), 341–350. https://doi.org/10.1016/j.adolescence.2012.12.001

Bauman, S., Toomey, R.B., & Walker, J.L. (2013) Associations among bullying, cyberbullying, and suicide in high school students. Journal of Adolescence 36(2): 341-50. https://doi.org/10.1016/j.adolescence.2012.12.001

Belsey, B. (2004) Cyberbullying: An Emerging Threat to the "Always On" Generation. https://www.cyberbullying.ca/pdf/Cyberbullying_Article_by_Bill_Belsey.pdf

Benjamin, R. (2019). Race after technology: Abolitionist tools for the New Jim Code. Polity.

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. European Journal of Communication, 33(2), 122–139. https://doi.org/10.1177/0267323118760317

Berlant, L. (2022). On the inconvenience of other people. Duke University Press.

Binny, M., Illendula, A., Saha, P., Sarkar, S., Goyal, P., & Mukherjee, A. (2020). Hate begets hate: A temporal study of hate speech. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2), 1–24. https://doi.org/10.1145/3415207

Blackwell, L. (2025). Content Moderation Futures. arXiv. https://arxiv.org/abs/2509.09076

Bronfenbrenner, U. (1979). The ecology of human development experiments by nature and design. Harvard University Press.

Buil-Gil, D., Kemp, S., Kuenzel, S., Coventry, L., Zakhary, S., Tilley, D., & Nicholson, J. (2023). The digital harms of smart home devices: A systematic literature review. Computers in Human Behavior, 145, 107770. https://doi.org/10.1016/j.chb.2023.107770

Butler, J. (2020). The force of nonviolence: An ethico-political bind. Verso.

Cachopo, J. P. (2022). The digital pandemic: Imagination in times of isolation (R. McGill, Trans.). Bloomsbury.

Carlson, C.R. (2021). Hate Speech. The MIT Press.

Cefai, S. (2020). Humiliation's media cultures: On the power of the social to oblige us. New Media & Society, 22(7), 1287–1304. https://doi.org/10.1177/1461444819879890

Cheng, J., Danescu-Nucuulescu-Mizil, C., & Leskovee, J. (2015). Antisocial Behavior in Online Discussion Communities. Ninth International AAAI Conference on Web and Social Media. https://cs.stanford.edu/people/jure/pubs/trolls-icwsm15.pdf

Christchurch Call (2019) Christchurch Call to Eliminate Violent and Terrorist Content.: https://www.christchurchcall.com

Cinelli, M., Pelicon, A., Mozetič, I., Quattrociocchi, W., Novak, P. K., & Zollo, F. (2021). Dynamics of online hate and misinformation. Scientific Reports, 11, 22083. https://doi.org/10.1038/s41598-021-01235-1

Costello, M., Rukus, J., & Hawdon, J. (2019). We don't like your type around here: Regional and residential differences in exposure to online hate material targeting sexuality. Deviant Behavior, 40(3), 385–401. https://doi.org/10.1080/01639625.2018.1426266

Cover, R. (2004). Digital Addiction: The Cultural Production of Online and Video Game Junkies. Media International Australia 113(1): 110-123. https://doi.org/10.1177/1329878X0411300113

Cover, R. (2022). Deepfake culture: The emergence of audio-video deception as an object of social anxiety and regulation. Continuum, 36(4), 609–621. https://doi.org/10.1080/10304312.2022.2090501

Cover, R. (2025). AI generation of rage bait: Implications for digital harms. New Media & Society. https://doi.org/10.1177/14614448251400675

Cover, R., Beckett, J., Brevini, B., Lumby, C., Simcock, R., & Thompson, J.D. (2025). Reporting online abuse to platforms: Factors, interfaces and the potential for care. Convergence: The International Journal of Research into New Technologies 32(1): 142-158. https://doi.org/10.1177/13548565251324508

Cover, R., Henry, N., Gleave, J., Greenfield, S., Grechyn, V., & Huynh, T.B. (2025). Protecting public figures online: How do platforms and regulators define public figures? Media International Australia 196(1): 156-170.

Cover, R., Humphries, J., Richardson, I., & Harris, D. X. (2025). Restricting young people from digital platforms: Demographics and the experience of digital harms in the support of social media age bans. Continuum: Journal of Media & Cultural Studies. https://doi.org/10.1080/10304312.2025.2598632

Cover, R., Simcock, R., & Humphries, J. (2025). Digital harms and penalties: Australian regulation, platform moderation and the figure of the perpetrator. Media International Australia. https://doi.org/10.1177/1329878X

Cross, K. (2019). Toward a formal sociology of online harassment. Human Technology, 15(3): 326–346. https://doi.org/10.17011/ht/urn.201911265023

Digital Action. (2023). Digital Action's online harms taxonomy. https://digitalaction.co/wp-content/uploads/2023/07/Digital-Actions-online-harms-taxonomy-1.pdf

Duff, C. (2026). The affective organisation of social infrastructures. Social Science & Medicine 390(1), 118873. https://doi.org/10.1016/j.socscimed.2025.118873

Engels, J. (2015). The Politics of Resentment: A Genealogy. The Pennsylvania State University Press.

Enock, F., Johansson, P., Bright, J., & Margetts, H. (2023) Tracking Experiences of Online Harms and Attitudes Towards Online Safety Interventions. Turing Institute. https://www.turing.ac.uk/sites/default/files/2023-03/tracking_experiences_of_online_harms_and_attitudes_report_final.pdf

Eribon, D. (2004). Insult and the Gay Self, trans. M. Lucey. Duke University Press.

eSafety Commissioner. (2020). Online Hate Speech. www.esafety.gov.au/sites/default/files/2020-01/Hate%20speech-Report.pdf.

eSafety Commissioner. (2022). Australians' negative online experiences 2022. https://www.esafety.gov.au/research/australians-negative-online-experiences-2022

eSafety Commissioner. (2024). Submission to the Joint Select Committee on Social Media and Australian Society, 21 June. https://www.esafety.gov.au/sites/default/files/2024-08/eSafety-Commissioner-submission-to-the-Joint-Select-Committee-on-Social-Media-and-Australian-Society.pdf?v=1724805626628

eSafety Commissioner. (2025). Hate in the Digital Age: Adults' Encounters with Online Hate. Canberra: Australian Government. https://www.esafety.gov.au/sites/default/files/2025-04/Hate-in-the-digital-age-adults-encounters-with-online-hate.pdf

European Commission. (2024). Standard Eurobarometer 101: Public opinion in the European Union, Spring. https://europa.eu/eurobarometer

European Commission. (2025). Cyberbullying: a common, EU-wide approach could help design effective response. The Joint Research Centre: EU Science Hub, 4 December. https://joint-research-centre.ec.europa.eu/jrc-news-and-updates/cyberbullying-common-eu-wide-approach-could-help-design-effective-response-2025-12-04_en

Fallis, D. (2015). What is disinformation? Library Trends, 63(3), 401–426. https://doi.org/10.1353/lib.2015.0009

Fleming, M.J.;, Greentree, S., Cocotti-Muller, D., Elias, K.A., & Morrison, S. (2006). Safety in cyberspace: Adolescents' safety and exposure online. Youth & Society 38(2): 135-154. https://doi.org/10.1177/0044118X06287858

Flew, T. (2021). Regulating platforms. Polity.

Flynn, A., Powell, A., Eaton, A., & Scott, A. J. (2025). Sexualized deepfake abuse: Perpetrator and victim perspectives on the motivations and forms of non-consensually created and shared sexualized deepfake imagery. Journal of Interpersonal Violence. Advance online publication. https://doi.org/10.1177/08862605241234567

Flynn, A., Vakhitova, Z., Wheildon, L., Harris, B., & Robards, B. (2025). Content moderation and community standards: The disconnect between policy and user experiences reporting harmful and offensive content on social media. Policy & Internet 17(3): e70006. https://doi.org/10.1002/poi3.70006

Franks, M. A. (2019). The cult of the Constitution. Stanford University Press.

Geoghegan, M. (2023). Online Harassment Negatively Impacts Users. Maryam Publishers.

George, C. (2016). Hate Spin: The Manufacture of Religious Offense and Its Threat to Democracy. The MIT Press.

Gillespie, T. (2018). Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), Article 2053951719897945. https://doi.org/10.1177/2053951719897945.

Grohol, J.M. (2000). Review: Caught in the Net. Addiction, 95(1): 139-140.

Haidt, J. (2024). The Anxious Generation: How the Great Rewiring of Childhood Is Causing an Epidemic of Mental Illness. Penguin.

Hannan, J. (2024). Trolling ourselves to Death: Democracy in the Age of Social Media. Oxford University Press.

Haraway, D. (1990). Simians, cyborgs, and women: The reinvention of nature. Routledge.

Harvey, D. (2017). Case note: Police v B [2017] NZHC 526, [2017] 3 NZLR 203. New Zealand Criminal Law Review. https://www.nzlii.org/nz/journals/NZCrimLawRw/2017/13.html

Harwell, D. (2026). White House posts altered image showing arrested protester crying. The Age, 23 January. https://www.theage.com.au/world/north-america/white-house-edits-crying-face-onto-photo-of-anti-ice-activist-s-arrest-20260123-p5nwg8.html

Haslop, C., O'Rourke, F., & Southern, R. (2021). #NoSnowflakes: The toleration of harassment and an emergent gender-related digital divide in a UK student online culture. Convergence, 27(5), 1418–1438. https://doi.org/10.1177/13548565211024891

Hayes, A.S. (2017). Sympathy for the Cyberbully: How the Crusade to Censor Hostile and Offensive Online Speech Abuses Freedom of Expression. Peter Lang.

Hayes, C. (2025). The Sirens' Call: How Attention Became the World's Most Endangered Resource. Scribe.

Heng, J., Bernstein, M., Danescu-Nuculescu-Mizil, C., & Leskovee, J. (2017). Anyone can become a troll: Causs of trolling behavior in online discussions. In: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing, 1217-1230. https://www.cs.cornell.edu/

Henry, N., Flynn, A., & Powell, A. (2020). Technology-facilitated domestic and sexual violence: A review. Violence Against Women, 26(15–16), 1828–1854. https://doi.org/10.1177/1077801219875821

Hettiachchi, D., Ji, K., Kennedy, J., McCosker,

A., Salim, F.D., Sanderson, M., Scholer, F., & Spina, D. (2023). Designing and evaluating presentation strategies for fact-checked content. Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 751–761. https://doi.org/10.1145/3583780.3614841

Hilliard, J., & Bhatt, A. (2025). Social Media Addiction. Addiction Center, 28 July. https://www.addictioncenter.com/behavioral-addictions/social-media-addiction/

Hindman, M. (2018). The Internet Trap: How the Digital Economy Builds Mobopolis and Undermines Democracy. Princeton University Press.

Hinduja, S. & Patchin, J.W. (2010). Bullying, cyberbullying, and suicide. Archives of Suicide Research 14(3): 206-221. https://doi.org/10.1080/13811118.2010.494133

Hinduja, S., & Patchin, J. W. (2014). Cyberbullying: Identification, prevention, and response. Cyberbullying Research Center. https://cyberbullying.org/Cyberbullying-Identification-Prevention-Response.pdf

Holly, L. (2024). Tackling digital harms: Why simply banning children from social media won't protect them. BMJ, 387, q2617. https://doi.org/10.1136/bmj.q2617

Hurcombe, E., Dehghan, E., Vodden, L., & Angus, D. (2025). The discursive function of Meta's Newsroom: How Meta frames the problem of problematic online content. Convergence: The International Journal of Research into New Media Technologies, 31(5), 1649-1671. https://doi.org/10.1177/13548565251315521.

Jane, E. A. (2015). Flaming? What flaming? The pitfalls and potentials of researching online hostility. Ethics and Information Technology, 17(1), 65–87. https://doi.org/10.1007/s10676-015-9362-0

Jane, E. A. (2017). Misogyny online: A short (and brutish) history. Sage.

Jiang, J.A., Scheuerman, M.K., Fiesler, C., & Brubaker, J.R. (2021). Understanding international perceptions of the severity of harmful content online. Plos One 16(8): e2025762. https://doi.org/10.1371/journal.pone.0256762

Kaye, D. (2019a). Promotion and Protection of the Right to Freedom of Opinion and Expression: Note by the Secretary-General, United Nations General Assembly. https://digitallibrary.un.org/record/3833657?ln=en

Kaye, D. (2019b) Speech Police: The Global Struggle to Govern the Internet. Columbia Global Reports.

Keighley, R. (2022). Hate hurts: Exploring the impact of online hate on LGBTQ+ young people. Women & Criminal Justice, 32(1–2), 29–48. https://doi.org/10.1080/08974454.2021.1988034

Kelly, O. (2023). New horizons or business as usual? New Zealand's medico-legal response to digital harm. Laws, 12(2), Article 12. https://doi.org/10.3390/laws12020032

Kirkland, T. (2025). Agree to disagree: Trust and civil debate. In: T. Kirkland and G. Fang (eds.), Age of Doubt: Building Trust in a World of Misinformation, 41-53. Monash University Publishing.

Koo, G. H., Masullo, G. M., Orr, B., & Huang, E. (2024). 'What flipping right does a teacher have to say being [LGBTQ+] is okay?' Understanding Twitter discourse around U.S. anti-LGBTQIA+ legislation. Howard Journal of Communications, 1–17. https://doi.org/10.1080/10646175.2024.2421859

Koukopoulos, N., Janickyj, M., & Tanczer, L. M. (2025). Defining and conceptualizing technology-facilitated abuse ("tech abuse"): Findings of a global Delphi study. Journal of Interpersonal Violence. Advance online publication. https://doi.org/10.1177/08862605241310465

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis. Psychological Bulletin, 140(4), 1073–1137. https://doi.org/10.1037/a0035618

KPMG and The University of Melbourne (2025). Trust, Attitudes and Use of Artificial Intelligence: A Global Study. https://kpmg.com

La Sala, L., Sabo, A., Michail, M., Thorn, P., Lamblin, M., Brown, V., & Robinson, J. (2024). Online safety when considering self-harm and suicide-related content: Qualitative focus group study with young people, policymakers, and social media industry professionals. Journal of Medical Internet Research. https://doi.org/10.2196/66321

LaRose, R., Kim, J. & Peng, W. (2011). Social networking: Addictive, compulsive, problematic, or just another media habit?' In: Z. Papacha-

rissi (ed.), A Networked Self: Identity, Community, and Culture on Social Network Sites, pp. 59-81. Routledge.

Levy, K., & Schneier, B. (2020). Privacy threats in intimate relationships. Journal of Cybersecurity, 6(1), tyaa006. https://doi.org/10.1093/cybsec/tyaa006

Marchal, N., et al. (2024). Generative AI misuse: A taxonomy of tactics and insights from real-world data. arXiv. https://doi.org/10.48550/arXiv.2406.13843

Mariconti, E., Suarez-Tangil, G., Blackburn, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Serrano, J. L., & Stringhini, G. (2019). 'You know what to do': Proactive detection of YouTube videos targeted by coordinated hate attacks. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW): 207. https://doi.org/10.1145/3359309

Marwick, A., & Lewis, R. (2016). Data and Society Report: Media Manipulation and Disinformation Online. https://datasociety.net/library/media-manipulation-and-disinfo-online/

Marwick, A., & Lewis, R. (2017). Media manipulation and disinformation online. Data & Society Research Institute.

Matsuda, M.J., Lawrence, C.R., Delgado, R., & Crenshaw, K.W. (2018). Words That Wound: Critical Race Theory, Assaultive Speech, And The First Amendment. Westview Press.

McKay, D., & Miller, C. (2021). Standing in the way of control: A call to action to prevent abuse through better design of smart technologies. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 302(1). https://doi.org/10.1145/3411764.3445114

McMahon, C. (2014). Why We Need a New Theory of Cyberbullying. http://dx.doi.org/10.2139/ssrn.2531796

McRobbie, A. (2020). Feminism and the politics of resilience: Essays on gender, media and the end of welfare. Polity.

Mead, M. (2017). Continuities in Cultural Evolution. Routledge.

Meikle, G. (2023). Deepfakes. Polity.

Meta. (2025). Bullying and harassment. Transparency Center. https://transparency.meta.com/en-gb/policies/community-standards/bullying-harassment/

Micallef, D., Parker, L., Brennan, L., Schivinski, B., & Jackson, M. (2022). Improving the health of emerging adult gamer: A scoping review of influences. Nutrients, 14(11). https://doi.org/10.3390/nu14112226

Micallef, D., Schivinski, B., Brennan, L., Parker, L., & Jackson, M. (2024). 'What are you eating?' Is the Influence of Fortnite streamers expanding beyond the game? Journal of Electronic Gaming and Esports, 2(1), 2023-0033. https://doi.org/10.1123/jege.2023-0033.

Mickelberg, A., Walker, B., Ecker, U., Howe, P.D.L., Perfors, A., & Fay, N. (2024). Does mud really stick? No evidence for continued influence of misinformation on newly formed person impressions. Collabra: Psychology, 10(1): 92332. https://doi.org/10.1525/collabra.92332

Mitts, T. (2025). Safe Havens for Hate: The Challenge of Moderating Online Extremism. Princeton University Press.

Moghaddam, S. H., et al. (2025). Towards a comprehensive taxonomy of online abusive language informed by machine learning. arXiv. https://arxiv.org/abs/2504.17653

Mosco, V. (2017). Becoming digital: Toward a post-internet society. Emerald.

Moseley, A. (2024). What to do if you've been doxxed. ABC, 15 May. https://www.abc.net.au/btn/high/what-to-do-if-you-ve-been-doxxed/103847288

Nash, V., & Felton, L. (2024). Treating the symptoms or the disease? Analysing the UK Online Safety Act's approach to digital regulation. Policy & Internet, 16(1), 818–832. https://doi.org/10.1002/poi3.404

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.

Noman, N. 2025. The new Meta policy makes social media a dangerous place for people like me. MSNBC, 16 January. https://www.msnbc.com/opinion/msnbc-opinion/meta-zuckerberg-trans-lgbtq-hate-speech-rcna187696.

Ofcom (2024). Deepfakes: Demean, Defraud and Disinform. https://www.ofcom.org.uk.

Orton-Johnson, K. (2024). Digital culture and society. Sage.

Phillips, W. (2015). This is Why We Can't Have Nice Things: The Relationship Between Onli-

ne Trolling and Mainstream Culture. MIT Press.

Phillips, W. (2018). The Oxygen of Amplification: Better Practices for Reporting on Extremists, Antagonists, and Manipulators online. Data & Society Research Institute.

Phillips, L. J., Pathé, M., & McEwan, T. (2023). Gender differences in stalking, threats and online abuse reported by Victorian politicians. Psychiatry, Psychology and Law, 30(6), 909–930. https://doi.org/10.1080/13218719.2022.2142975

Poland, B. (2016). Haters: Harassment, abuse, and violence online. University of Nebraska Press.

Pond, P. (2024). Asynchronicity: The Temporal Dimensions of the Information Crisis. Walter de Gruyter.

Radio Free Asia. (2022). The Chinese Internet's Hidden Victims: Uncovering and healing the scars of online abuse. https://www.wainao.me/wainao-reads/uncovering-and-healing-the-scars-of-online-abuse-04132022

Reddit. (2025). Do not threaten, harass, or bully. Reddit Help. https://support.reddithelp.com/hc/en-us/articles/360043071072-Do-not-threaten-harass-or-bully

Salter, A., & Blodgett, B. (2017). Toxic geek masculinity in media: Sexism, trolling, and identity policing. Palgrave Macmillan.

Schoenebeck, S., Lampe, C., & Triêu, P. (2023). Online harassment: Assessing harms and remedies. Social Media + Society. https://doi.org/10.1177/20563051231157297

Shackleton, N. (2024). The government is drafting anti-hate speech laws. Here are 4 things they should include. The Conversation, 12 June. https://theconversation.com/the-government-is-drafting-anti-hate-speech-laws-here-are-4-things-they-should-include-231178

Shelby, R., et al. (2023). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. HAL. https://doi.org/10.1145/3600211.3604673

Simon, H.A. (1971). Designing organizations for an information-rich world. In M. Greenberger (ed.), Computers, Communications, and the Public Interest, pp. 38-72. The Johns Hopkins Press.

Slupska, J. (2019). Safe at home: Towards a feminist critique of cybersecurity. St Antony's International Review, 15(1), 83-100. https://www.jstor.org/stable/27027755

Stokel-Walker, C. (2024). Politicians say they can make social media less 'addictive'. Experts aren't so sure. BBC, 18 September. https://www.bbc.com/future/article/20240626-can-a-law-make-social-media-less-addictive

Strossen, N. (2018). Hate: Why We Should Resist It with Free Speech, Not Censorship. Oxford University Press.

Suzor, N. P. (2019). Lawless: The secret rules that govern our digital lives. Cambridge University Press.

Terranova, T. (2022). After the internet: Digital networks between capital and the common. Semiotext(e).

Thomas, C., Cojocaru, M., & Rosenberg, N. (2022). The Hate Speech Crisis: Ways to start fixing it – A toolkit for civil society organizations and activists. Minority Rights Group International.

Thompson, E.P. (1967). Time, work-discipline, and industrial capitalism. Past & Present 38(Dec): 56-97. https://doi.org/10.1093/past/38.1.56

Thompson, J. D., & Cover, R. (2021). Digital hostility, internet pile-ons and shaming: A case study. Convergence, 28(6), 1770–1782. https://doi.org/10.1177/13548565211030461

Thompson, S. A., & Conger, K. (2025, January 7). Meet the next fact-checker, debunker and moderator: You. The New York Times. https://www.nytimes.com/2025/01/07/technology/meta-facebook-content-moderation.html

TikTok. (2025a). Harassment and bullying. Safety and Civility. https://www.tiktok.com/community-guidelines/en/safety-civility?cgversion=2025H2update#7

TikTok. (2025b). Bullying prevention. Safety Center. https://www.tiktok.com/safety/en/bullying-prevention/

Tobi, A. (2024). Towards an Epistemic Compass for Online Content Moderation. Philosophy and Technology 37(3): 1–20. https://doi.org/10.1007/s13347-024-00791-3

Tong, S.T. (2025). Foundations, definitions, and directions in online hate research. In: J.B. Walther and R.E. Rice (eds.), Social Processes of Online Hate, 37-72. Routledge.

Törnberg, P. (2022). How digital media drive affec-

tive polarization through partisan sorting. Proceedings of the National Academy of Sciences, 119(42), e2207159119. https://doi.org/doi:10.1073/pnas.2207159119

Unger, M. (2012). Social ecologies and their contribution to resilience. In M. Unger (Ed.), The social ecology of resilience: A handbook of theory and practice (pp. 13–31). Springer.

United Nations. (2025). Understanding hate speech. https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

Villar Onrubia, D., Barreda Angeles, M., Cachia, R., Economou, A., & Lopez Cobo, M., (2025). Cyberbullying: Insights from Science, Policy and Legislation. Publications Office of the European Union. https://data.europa.eu/doi/10.2760/0941861.

Vincent, N. A. (2017). Victims of cybercrime: Definitions and challenges. In E. Martellozzo & E. A. Jane (Eds.), Cybercrime and its victims (pp. 27–42). Routledge.

Vogels, E.A. (2021). The state of online harassment. Pew Research Center, 13 January. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/.

Wachs, S., & Wright, M. F. (2018). Associations between bystanders and perpetrators of online hate: The moderating role of toxic online disinhibition. International Journal of Environmental Research and Public Health, 15(9), 2030. https://doi.org/10.3390/ijerph15092030

Waldron, J. (2012). The harm in hate speech. Harvard University Press.

Waldron, J. (2012). The Harm in Hate Speech. Harvard University Press.

Walther, J.B. (2025). Making a case for a social processes approach to online hate. In: J.B. Walther and R.E. Rice (eds.), Social Processes of Online Hate, 9-36. Routledge.

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking. Council of Europe. https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html

Wijenayake, S., Gray, J., Jayatilaka, A., La Sala, L., Arachchilage, N.A.G., & Kelly, R.M. (2025). Advancing interdisciplinary approaches to online safety research. OZCHI 25: 37th Australian Conference on Human-Computer Interaction. https://dl.acm.org/doi/pdf/10.1145/3764687.3767275

World Economic Forum. (2023). Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms. https://www.weforum.org/publications/toolkit-for-digital-safety-design-interventions-and-innovations-typology-of-online-harms/

World Health Organization. (2002). World report on violence and health. https://www.who.int/publications/i/item/9241545615

Wu, T. (2025). The Age of Extraction: How Tech Platforms Conquered the Economy and Threaten our Future Prosperity. The Bodley Head.

X. (2024). Abuse and harassment. https://help.x.com/en/rules-and-policies/abusive-behavior

Xie, X., Cambazoglu, I., Berger-Correa, B., & Ringrose, J. (2022). Anti-feminist misogynist shitposting: The challenges of feminist academics navigating toxic twitter. In P.J. Burke, J. Coffey, R. Gill & A. Kanai (eds.), Gender in an Era of Post-truth Populism: Pedagogies, Challenges and Strategies. London: Bloomsbury.

Young, K.S. (1998). Caught in the Net: How to Recognize the Signs of Internet Addiction – and a Winning Strategy for Recovery. John Wiley & Sons.

Young, R., Subramanian, R., Miles, S., Hinnant, A., & Andsager, J. L. (2017). Social representation of cyberbullying and adolescent suicide: A mixed-method analysis of news stories. Health Communication, 32(9), 1082–1092. https://doi.org/10.1080/10410236.2016.1214214

YouTube. (2025). Harassment and cyberbullying policies. YouTube Help. https://support.google.com/youtube/answer/2802268

Zhou, X., Qin, D., Lu, X., Chen, L., & Zhang, Y. (2019). Online social media recommendation over streams. 2019 IEEE 35th International Conference on Data Engineering (ICDE), Macao, China, 938-949. https://doi.org/10.1109/ICDE.2019.00088.

Schoenebeck, S., Lampe, C., & Triêu, P. (2023). Online harassment: Assessing harms and remedies. Social Media + Society. https://doi.org/10.1177/20563051231157297

Shackleton, N. (2024). The government is drafting anti-hate speech laws. Here are 4 things they

should include. The Conversation, 12 June. https://theconversation.com/the-government-is-drafting-anti-hate-speech-laws-here-are-4-things-they-should-include-231178

Shelby, R., et al. (2023). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. HAL. https://doi.org/10.1145/3600211.3604673

Simon, H.A. (1971). Designing organizations for an information-rich world. In M. Greenberger (ed.), Computers, Communications, and the Public Interest, pp. 38-72. The Johns Hopkins Press.

Slupska, J. (2019). Safe at home: Towards a feminist critique of cybersecurity. St Antony's International Review, 15(1), 83-100. https://www.jstor.org/stable/27027755

Stokel-Walker, C. (2024). Politicians say they can make social media less 'addictive'. Experts aren't so sure. BBC, 18 September. https://www.bbc.com/future/article/20240626-can-a-law-make-social-media-less-addictive

Strossen, N. (2018). Hate: Why We Should Resist It with Free Speech, Not Censorship. Oxford University Press.

Suzor, N. P. (2019). Lawless: The secret rules that govern our digital lives. Cambridge University Press.

Terranova, T. (2022). After the internet: Digital networks between capital and the common. Semiotext(e).

Thomas, C., Cojocaru, M., & Rosenberg, N. (2022). The Hate Speech Crisis: Ways to start fixing it – A toolkit for civil society organizations and activists. Minority Rights Group International.

Thompson, E.P. (1967). Time, work-discipline, and industrial capitalism. Past & Present 38(Dec): 56-97. https://doi.org/10.1093/past/38.1.56

Thompson, J. D., & Cover, R. (2021). Digital hostility, internet pile-ons and shaming: A case study. Convergence, 28(6), 1770–1782. https://doi.org/10.1177/13548565211030461

Thompson, S. A., & Conger, K. (2025, January 7). Meet the next fact-checker, debunker and moderator: You. The New York Times. https://www.nytimes.com/2025/01/07/technology/meta-facebook-content-moderation.html

TikTok. (2025a). Harassment and bullying. Safety and Civility. https://www.tiktok.com/community-guidelines/en/safety-civility?cgversion=2025H2update#7

TikTok. (2025b). Bullying prevention. Safety Center. https://www.tiktok.com/safety/en/bullying-prevention/

Tobi, A. (2024). Towards an Epistemic Compass for Online Content Moderation. Philosophy and Technology 37(3): 1–20. https://doi.org/10.1007/s13347-024-00791-3

Tong, S.T. (2025). Foundations, definitions, and directions in online hate research. In: J.B. Walther and R.E. Rice (eds.), Social Processes of Online Hate, 37-72. Routledge.

Törnberg, P. (2022). How digital media drive affective polarization through partisan sorting. Proceedings of the National Academy of Sciences, 119(42), e2207159119. https://doi.org/doi:10.1073/pnas.2207159119

Unger, M. (2012). Social ecologies and their contribution to resilience. In M. Unger (Ed.), The social ecology of resilience: A handbook of theory and practice (pp. 13–31). Springer.

United Nations. (2025). Understanding hate speech. https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

Villar Onrubia, D., Barreda Angeles, M., Cachia, R., Economou, A., & Lopez Cobo, M., (2025). Cyberbullying: Insights from Science, Policy and Legislation. Publications Office of the European Union. https://data.europa.eu/doi/10.2760/0941861

Vincent, N. A. (2017). Victims of cybercrime: Definitions and challenges. In E. Martellozzo & E. A. Jane (Eds.), Cybercrime and its victims (pp. 27–42). Routledge.

Vogels, E.A. (2021). The state of online harassment. Pew Research Center, 13 January. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

Wachs, S., & Wright, M. F. (2018). Associations between bystanders and perpetrators of online hate: The moderating role of toxic online disinhibition. International Journal of Environmental Research and Public Health, 15(9), 2030. https://doi.org/10.3390/ijerph15092030

Waldron, J. (2012). The harm in hate speech. Harvard University Press.

Waldron, J. (2012). The Harm in Hate Speech. Har-

vard University Press.

Walther, J.B. (2025). Making a case for a social processes approach to online hate. In: J.B. Walther and R.E. Rice (eds.), Social Processes of Online Hate, 9-36. Routledge.

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking. Council of Europe. https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html

Wijenayake, S., Gray, J., Jayatilaka, A., La Sala, L., Arachchilage, N.A.G., & Kelly, R.M. (2025). Advancing interdisciplinary approaches to online safety research. OZCHI 25: 37th Australian Conference on Human-Computer Interaction. https://dl.acm.org/doi/pdf/10.1145/3764687.3767275

World Economic Forum. (2023). Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms. https://www.weforum.org/publications/toolkit-for-digital-safety-design-interventions-and-innovations-typology-of-online-harms/

World Health Organization. (2002). World report on violence and health. https://www.who.int/publications/i/item/9241545615

Wu, T. (2025). The Age of Extraction: How Tech Platforms Conquered the Economy and Threaten our Future Prosperity. The Bodley Head.

X. (2024). Abuse and harassment. https://help.x.com/en/rules-and-policies/abusive-behavior

Xie, X., Cambazoglu, I., Berger-Correa, B., & Ringrose, J. (2022). Anti-feminist misogynist shitposting: The challenges of feminist academics navigating toxic twitter. In P.J. Burke, J. Coffey, R. Gill & A. Kanai (eds.), Gender in an Era of Post-truth Populism: Pedagogies, Challenges and Strategies. London: Bloomsbury.

Young, K.S. (1998). Caught in the Net: How to Recognize the Signs of Internet Addiction – and a Winning Strategy for Recovery. John Wiley & Sons.

Young, R., Subramanian, R., Miles, S., Hinnant, A., & Andsager, J. L. (2017). Social representation of cyberbullying and adolescent suicide: A mixed-method analysis of news stories. Health Communication, 32(9), 1082–1092. https://doi.org/10.1080/10410236.2016.1214214

YouTube. (2025). Harassment and cyberbullying policies. YouTube Help. https://support.google.com/youtube/answer/2802268

Zhou, X., Qin, D., Lu, X., Chen, L., & Zhang, Y. (2019). Online social media recommendation over streams. 2019 IEEE 35th International Conference on Data Engineering (ICDE), Macao, China, 938-949. https://doi.org/10.1109/ICDE.2019.00088

RMIT Digital Ethnography Research Centre

124 La Trobe Street

Building 6, Level 4

Melbourne City Campus

digital-ethnography.com

E digital.ethnography@rmit.edu.au

L Follow us on LinkedIn