



April 2026

Judicial Handbook

for navigating online harms



Table of contents

1. INTRODUCTION.....	3
2. WHAT ARE ONLINE HARMS	4
Online harassment	5
Misinformation and disinformation	7
Data protection and privacy concerns	7
The collateral damage of content moderation	8
Hate speech and incitement	9
Algorithmic bias and discrimination	9
Amplification of extremism and polarisation	9
Financial cybercrime and fraud	10
3. EFFECTS OF ONLINE HARMS ON VICTIMS AND SURVIVORS	11
The rights implications of online harms	11
The psychological and social effects of online harms	14
The physical and offline effects of online harms	15
The particular effect of online harms on children	16
4. AN OVERVIEW OF THE MAIN SOCIAL MEDIA PLATFORMS.....	17
WhatsApp.....	17
X (previously Twitter)	19
Instagram.....	21
Facebook.....	24
TikTok.....	26
5. EXISTING LEGISLATION	29
The Cybercrimes Act and Electronic Communications and Transactions Act	29
The Protection of Personal Information Act.....	31
The Protection from Harassment Act	32
The Promotion of Equality and Prevention of Unfair Discrimination Act and the Prevention and Combating of Hate Crimes and Hate Speech Act.....	33
The Film and Publication Act	34
6. PRACTICAL CONSIDERATIONS	35
Patterns of harm	35
Context is key	36
Challenges of addressing online harms.....	39
The opportunities of technology.....	41
Evidence gathering and management	41
Experts	42
6. EMERGING TRENDS	43
8. CONCLUSION	45
ANNEXURE A: INTERNET USER GUIDE FOR BEGINNERS	46

1. Introduction

In today's interconnected digital world, online platforms offer immense opportunities for communication, expression, and information sharing. However, these spaces also harbour significant risks, exposing individuals to various forms of online harm. "Online harms" broadly refer to a range of detrimental behaviours and practices that occur in the digital sphere, negatively affecting individuals, communities, and society at large.¹ These harms can take numerous forms, each posing unique challenges to the well-being, privacy, and safety of those targeted. This handbook aims to provide a comprehensive understanding of online harms, focusing on key categories such as online harassment and data protection and privacy violations.

Understanding the scope and nuances of online harms is crucial, particularly for those tasked with adjudicating such cases, as they directly impact individuals' fundamental rights. These harms not only affect the victims and survivors psychologically, socially, and physically but also have far-reaching implications for their rights to dignity, equality, privacy, and safety, particularly for women, gender minorities, children and other vulnerable groups. With this foundation, the following sections will delve into the effects of online harms, the platforms where they commonly occur, existing legislation, and practical considerations for addressing them in judicial contexts. This will also include an exploration of emerging trends and opportunities technology offers to better address these issues, enabling a more informed and just approach to adjudicating online harms.

¹ World Economic Forum, 'Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms' (2023)(accessible [here](#)).



2. What are online harms

Online harms are multi-dimensional and often overlap, making it difficult to neatly categorise them. For example, the non-consensual distribution of intimate images (“NCII”) can violate privacy, dignity, and even endanger personal safety, while child sexual exploitation encompasses both physical and psychological violations. The World Economic Forum’s typology of online harms, which views these issues through a human rights lens, is particularly useful.² This approach highlights the impact on individuals and frames online harms in a way that is applicable across different jurisdictions, even while acknowledging that cultural and regional contexts affect how harms may manifest.

A study by the UK regulator, Ofcom, revealed that 62% of adults and 81% of internet users aged 12-15 experienced at least one harmful online incident each year.³ For adults, the most common issues were fraud, scams, and exposure to fake news.⁴ However, over a quarter of younger users reported being victims of online bullying and receiving unwanted friend requests. Nearly 40% of teenagers encountered offensive language, and almost a quarter had been subjected to trolling. Presently, there is a lack of a globally agreed definition of harmful content and online harms, and as a result, the World Economic Forum’s Global Coalition for Digital Safety developed a Typology of Online Harms, categorising harmful content into six distinct groups.⁵

² Jigsaw Research, ‘Internet users’ experience of potential online harms: summary of survey research’ (2020)(accessible [here](#)).

³ Id.

⁴ The use of the term “fake news” is discouraged due to its vagueness and use as an umbrella term for questionable information in general. See Ratcliffe “The term ‘fake news’ is doing great harm” *The Conversation* (27 July 2018)(accessible [here](#)).

⁵ See above n1.



The Global Coalition for Digital Safety has also introduced a high-level framework to help stakeholders understand and assess digital safety risks, assisting them in navigating new regulatory landscapes.⁶ The Coalition acknowledges that while the internet is global, the harms encountered can be very local or context-specific, with unique risks emerging in different countries, regions, and communities.⁷ In the South African context, it's crucial to recognise that ensuring digital safety is a complex process that involves legal, policy, ethical, social, and technological considerations. Decision-making in this regard must be firmly grounded in international human rights frameworks.

The primary categories of online harm discussed in this section include online harassment, misinformation and disinformation, data protection and privacy breaches, the potential consequences of content moderation, hate speech and incitement, algorithmic bias and discrimination, extremism and polarisation, and financial cybercrime and fraud.

Online harassment

Online harassment severely restricts individuals' ability to exercise their rights, particularly the rights to privacy, freedom of expression, and safety. This harassment manifests in various forms, including cyberbullying, cyberstalking, hate speech, NCII, and trolling. These behaviours aim to intimidate, discredit, silence, or shame their targets, often disproportionately affecting marginalised groups like women, racial minorities, and LGBTQ+ individuals.⁸ Platforms such as Facebook, Instagram, and X (formerly Twitter) are commonly used to perpetuate such harassment, though these behaviours can occur on any online platform.⁹

The consequences of online harassment often extend beyond the digital world. Victims and survivors may suffer severe psychological distress, fear for their physical safety, or face irreparable damage to their reputations.¹⁰ In some cases, individuals may be forced to censor their views or withdraw from online spaces entirely, thereby stifling their freedom of expression. For instance, women who are disproportionately targeted by sexualised harassment may feel compelled to self-censor, leaving important public conversations dominated by a less diverse group of voices.¹¹ A particularly harmful subset of online harassment is NCII. This form of harassment severely violates the victim's privacy, often leading to public shaming, extortion, and lasting psychological trauma. Legal frameworks, such as recent laws in South Africa, now criminalise these acts but enforcement remains a challenge.¹² A recent example of a case in South Africa about the non-consensual sharing of intimate images will be canvassed below:¹³

6 World Economic Forum, 'How to measure digital safety effectively to reduce risks online' (2024)([accessible here](#)).

7 Id.

8 United Nations Women, 'Online and ICT facilitated violence against women and girls during COVID-19' (2020)([accessible here](#)).

9 Id.

10 Pew Research Center, 'The State of Online Harassment' (2021)([accessible here](#)).

11 Id.

12 Cybercrimes Act 19 of 2020 at section 16.

13 AfriForum, 'Encouraging development in so-called revenge porn case after the accused is arrested and appears in court' (2023)([accessible here](#)); L Mashigo 'AfriForum tackles 'revenge porn' abuse case' (2023)([accessible here](#)); and T Head, 'Pretoria's worst bully could be brought down by revenge porn case' (2022)([accessible here](#)).





Private prosecution against NCII

A recent case in South Africa highlights the issue of online harassment through NCII. Lindrie Gouws, a woman from Pretoria, accused her ex-partner, Johnny Rockett, of distributing her intimate photos to friends without her consent, with the intent to cause emotional distress. Gouws was supported by AfriForum's private prosecution unit, which helped her open a case in terms of the Film and Publications Act, the Cybercrimes Act, as well as *crimen injuria*. The accused was arrested but subsequently released on bail.

The case of Lindrie Gouws illustrates the harmful impact of NCII, often incorrectly labelled as "revenge porn". This term is increasingly recognised as problematic and misleading. Here is why:¹⁴



Why words matter: revenge porn versus NCII

Activists and researchers advocate for the use of more accurate terminology, such as NCII, to properly reflect the harm and remove the implied justification for the perpetrator's actions. Terminology is crucial in addressing online harms, particularly those that disproportionately affect marginalised groups. The term "revenge porn" suggests that the victim has committed an offense deserving of retaliation, reinforcing victim-blaming narratives.

Harassment of journalists, activists, and media professionals, particularly women, has surged in recent years. This not only infringes on the rights of these individuals but also threatens broader societal interests by undermining the free flow of information and creating a climate of fear.



¹⁴ S Maddocks, 'Revenge Porn: 5 important reasons why we should not call it by that name' (2019)(accessible [here](#)).



Misinformation and disinformation



Why words matter: Mis and disinformation not “fake news”¹⁵

The term “fake news” has gained popularity, driven largely by Donald Trump in the run-up to his first presidency. While it has recently become a popular term, people spreading lies dressed as news or using dramatic headlines to sell newspapers are old manipulation strategies. What is new, is our digital reality and social media where it is easy to invent, spread and promote “fake news”. The term “fake news” is often used to discredit credible news reporting and journalism it has become an effective means to undermine the credibility of all news media – whether they produce credible news or not. It is more accurate to use terms like mis and disinformation.

The spread of misinformation and disinformation presents another significant threat to users’ rights in the online space. While both involve the dissemination of false information, misinformation refers to the unintentional spread of inaccuracies, while disinformation is a deliberate attempt to deceive or manipulate. Both phenomena undermine informed decision-making and can endanger public safety, particularly during crises such as elections or public health emergencies.¹⁶

Disinformation campaigns, which are often coordinated, well-funded, and technologically sophisticated, pose a greater risk than misinformation due to their strategic intent. For example, disinformation campaigns surrounding elections or public health measures can severely impact democratic processes and public trust. Vulnerable populations, such as children, are particularly susceptible to these harms, as they may lack the critical thinking skills necessary to discern truth from falsehood in online content.¹⁷

Data protection and privacy concerns

Social media platforms collect vast amounts of data from their users, often without the user’s full understanding or consent. This collection includes not only the data users willingly provide such as posts, photos, and location data but also more covert data such as interaction metrics. This includes what content users view, share, or engage with.

¹⁵ See n4 above.

¹⁶ Media Monitoring Africa, ‘Discussion Document: Disinformation through a children’s rights lens’ (2022)(accessible [here](#)).

¹⁷ European Union, ‘Online Platforms, the digital single market and disinformation’ (2018)(accessible [here](#)).



Several key online harms related to data protection and privacy arise from this data collection:¹⁸

1. **Data breaches:** Social media platforms are frequent targets for cyberattacks, due to the vast amounts of personal information they store. When data breaches occur, users' sensitive information can be exposed, leading to identity theft, financial fraud, or other malicious uses.
2. **Third party data sharing:** Many platforms share user data with advertisers or third-party developers, often without clear consent from users. While privacy settings exist, they are frequently complex and insufficiently transparent, leaving users unaware of how their data is being used or who has access to it.
3. **Data mining and profiling:** Social media companies use algorithms to analyse user data and build detailed profiles for targeted advertising. While this can enhance user experience, it can also create echo chambers and filter bubbles, reinforcing biases and limiting exposure to diverse viewpoints. These practices also raise concerns about manipulation in political campaigns and other sensitive contexts.
4. **Surveillance and monitoring:** Social media platforms can be used for surveillance by governments and intelligence agencies. This surveillance often lacks transparency or judicial oversight, creating a chilling effect on free expression and privacy rights.

The collateral damage of content moderation

Content moderation practices, designed to prevent harm on social media platforms, carry their own risks. Moderation policies are often guided by vague standards, granting platforms significant discretion in deciding what content should be removed, amplified, or restricted.¹⁹ While these practices aim to curb harmful behaviour, they can also result in arbitrary censorship, undermining users' freedom of expression.

Moreover, social media companies have a financial incentive to promote content that generates engagement, which can often be inflammatory or emotionally charged.²⁰ This dynamic creates tension between protecting users from harm and maximising profit. In some cases, harmful content is not removed swiftly enough, while in others, legitimate content is censored unnecessarily.

Artificial Intelligence ("AI") driven content moderation adds another layer of complexity. Although AI systems are capable of processing vast amounts of content, they can also make biased or erroneous decisions, removing content pre-emptively or disproportionately targeting certain groups.²¹ This lack of transparency in AI decision-making raises concerns about discriminatory practices and the stifling of marginalised voices.

¹⁸ Verizon, 'Data Breach Investigations Report' (2024)(accessible [here](#)).

¹⁹ Legal Resources Centre, 'A critical analysis of content moderation policies and the impact of spreading violence, hatred and disinformation in the global south' (accessible [here](#)).

²⁰ Id.

²¹ J Canda 'The role of AI in content moderation and online safety' (2024)(accessible [here](#)).



Hate speech and incitement

According to the South African Human Rights Commission, hate speech is expression that goes beyond mere insults or offensive language – it infringes on the dignity of certain persons or groups.²² Hate speech is detrimental in that it negatively impacts on victims' self-worth, and causes significant hurt, harm, pain, distress and humiliation. Further, it impugns the right to human dignity and equality.

Incitement occurs when a person encourages or urges others to commit acts of violence or hatred. When assessing incitement, it is not necessary that the action that was being incited actually has to occur.²³ Hate speech and incitement are linked in that the effects thereof are similar. Incitement may infringe on the targeted group or person's dignity and equality. However, incitement may also lead to physical harm. Both harms may occur online, and are prohibited by South African law, which will be set out below.

Algorithmic bias and discrimination

Algorithms are central to the operation of social media platforms, determining what content users see, which posts are amplified, and which are suppressed.²⁴ While these algorithms are designed to enhance user experience, they can unintentionally perpetuate bias and discrimination. Algorithmic bias arises when systems are trained on biased data or when their design reflects the biases of their creators. This results in discriminatory outcomes, reinforcing stereotypes, promoting harmful content, or disproportionately targeting certain communities.²⁵

For example, studies have shown that algorithms often over-police the speech of minority groups while failing to detect hate speech directed at them. Women, racial minorities, and LGBTQ+ individuals are more likely to have their content flagged or removed, even when they are using the platform to express themselves or report abuse.²⁶ Such bias can contribute to the marginalisation of already vulnerable groups and stifle diverse voices online.

Amplification of extremism and polarisation

One of the most concerning harms facilitated by social media is the amplification of extremist content and the polarisation of public discourse. Social media platforms, through algorithms that prioritise engagement, tend to promote content that is inflammatory, controversial, or emotionally charged. This has led to the rapid spread of extremist ideologies, including xenophobia, white supremacy, religious fundamentalism, and political radicalism, which can incite violence and hatred both online and offline.

22 South African Human Rights Commission, 'Hate Speech' (accessible [here](#)).

23 Southern Africa Litigation Centre, 'Hate Speech and Incitement' (accessible [here](#)).

24 T Coleman 'How content creators cope with discriminatory algorithms' (2023) (accessible [here](#)).

25 Public Citizen 'Report: Algorithms are worsening racism, bias, discrimination' (2021) (accessible [here](#)).

26 Amnesty International 'Toxic Twitter- a toxic place for women' (2018) (accessible [here](#)).



The echo chamber effect, where users are primarily exposed to content that reinforces their existing beliefs, further exacerbates polarisation.²⁷ As individuals become more entrenched in their views, societal cohesion deteriorates, making it difficult to engage in meaningful, productive discourse. This polarisation has real-world consequences, such as increased political violence and the erosion of democratic norms. See below a recent case in South Africa, which illustrates how the amplification of xenophobia online has caused harm online and offline:²⁸



Protection order obtained for online harassment

In 2024, Nicole Fritz, the former executive director of the Helen Suzman Foundation (“HSF”), was granted a protection order by the Johannesburg Central Magistrates Court after facing harassment and abuse on social media. The abuse came after the HSF challenged the government’s decision to terminate Zimbabwean Exemption Permits, which had been in place since 2009 to allow Zimbabweans to live, work, and study in South Africa. The harassment, primarily from one individual on platform X, escalated to the point where Fritz sought legal protection. Fritz noted that other women had also been targeted by the same individual.

Financial cybercrime and fraud

Social media platforms have become fertile ground for various forms of financial cybercrime, including fraud, identity theft, phishing, and hacking. Cybercriminals exploit the openness and interconnectedness of social media platforms to target users with malicious schemes, often tricking them into sharing personal information or clicking on harmful links.²⁹ This poses significant risks to users’ financial security and personal privacy. One particularly harmful form of cybercrime is online scams, where fraudsters pose as legitimate businesses, charities, or even friends to steal money or data.³⁰ These scams can range from phishing emails to complex schemes that exploit vulnerable individuals, such as the elderly or those in financial distress.

Online harms are complex and multifaceted, affecting individuals’ enjoyment of their constitutional rights and freedoms. While social media platforms play a central role in our digital lives, they have also become sites of significant risk.³¹ It is essential to address these harms with a nuanced understanding of their impact on vulnerable groups and to ensure that both the prevention of harm and the protection of rights are carefully balanced.³² For judges, this requires not only familiarity with existing legal frameworks but also a deep understanding of the platforms where these harms occur and the broader social contexts in which they manifest.

27 P Barbera ‘Social Media, Echo Chambers, and Political Polarization’ (2020) ([accessible here](#)).

28 T Pongweni ‘Former Helen Suzman Foundation chief granted court order to halt cyberbully after ZEP challenges’ (2024) ([accessible here](#)).

29 United Nations Office on Drugs and Crime ‘Cybercrime University Module Series’ ([accessible here](#)).

30 Id.

31 Id.

32 Id.



3.

Effects of online harms on victims and survivors

Online harms affect the victims and survivors thereof in a myriad of ways. This section will describe these effects. First, the implications of online harms on victims and survivors' rights will be discussed. Further, the psychological, physical and social consequences of online harms will be set out. The offline manifestations of online harm will also be explained. Last, the particular effects of online harms on children will be discussed.

The rights implications of online harms

Online harms infringe on the constitutional rights of victims and survivors. The extent of infringement, as well as which rights are affected, is dependent on the type of harm as well as the degree of harm. Further, a victim or survivor's rights may be affected differently from another who has experienced the same type of harm, depending on that person's particular context. The concept of intersectionality is important to understand in this regard.





Intersectionality

The concept of intersectionality states that different social identities, for example race, gender and sexuality, intersect and inform each other resulting in each individual experiencing harms uniquely.³³ For example, a black women's rights may be differently implicated than that of a white women. In *Mahlangu and Another v Minister of Labour and Others*,³⁴ the Constitutional Court highlighted and recognised the concept of intersectionality in assessing discrimination under section 9(3) of the Constitution. Here, the Court acknowledged that rights infringements may impact an individual in a "multiplicity of ways based on their position in society and the structural dynamics at play",³⁵ and stated that this position has been accepted in our law.³⁶

Intersectionality should be kept in mind when assessing the impact of online harms on rights, as the infringements may manifest differently depending on a person's particular characteristics and circumstances. The particular vulnerability of women and gender minorities, as well as intersectional considerations of race, ethnicity, disability, and nationality should be considered.

Online harms may infringe of a variety of constitutional rights. The specific rights affected will differ from case to case, in the same way that the effects of a physical harm are context specific. The below listed rights are often infringed on by online harms. It is not a closed list.

- **Dignity:** The Constitution provides everyone with the right to have their human dignity respected and protected.³⁷ Online harms may infringe peoples' dignity by attacking their self-worth, privacy, or reputation, and by causing psychological distress. Certain online harms may humiliate, shame or stigmatize the victim or survivor, thereby infringing on their right to dignity.³⁸ The public nature of online platforms means that harmful content can be widely seen and shared, which may amplify this impact.
- **Privacy:** Depending on the nature of the harm in question, online harms may infringe on peoples' right to privacy.³⁹ In *Botha v Smuts and Another*, the Constitutional Court found that the sharing of Botha's home address on a Facebook post violated his constitutional right to privacy.⁴⁰ Kollapen J, writing for the majority, found that Botha had a reasonable expectation of privacy over his home address even though the address was, prior to Smuts positing it on Facebook, publicly available. Botha's subjective expectation of privacy "may well exist in that a post relating to...his home address, even one widely published, could be a private fact".⁴¹ This

33 T Meer and A Müller, 'Considering intersectionality in Africa' (2017) at pg 3 (accessible [here](#)).

34 *Mahlangu and Another v Minister of Labour and Others* 2021(2) SA 54 (CC) (accessible [here](#)).

35 Id at para 76.

36 Earlier jurisprudence has recognised the concept of intersectionality, albeit implicitly. See, for example, *National Coalition for Gay and Lesbian Equality and Another v Minister of Justice and Others* 1999(1) SA 6 (CC) at para 113 (accessible [here](#)).

37 The Constitution of the Republic of South Africa, 1996 at section 10.

38 Report of the Special Rapporteur on Violence Against Women, Its Causes and Consequences on Online Violence Against Women and Girls from a Human Rights Perspective A/HRC/38/47 (2018) at para 57 (accessible [here](#)).

39 Id at section 14.

40 *Botha v Smuts and Another* (CCT 40/22) [2024] ZACC (accessible [here](#)).

41 Id at paras 135-136.



case serves as an example of where an online harm infringed on the right to privacy in that a person's private information in the form of their home address was exposed. NCDII serves as another example of where a person's right to privacy may be infringed. Our Courts have held that the non-consensual distribution of intimate images and the non-consensual possession thereof infringes on the victim or survivor's rights to privacy, sexual dignity, and bodily and psychological integrity.⁴²

- **Freedom and security of the person:** Section 12 of the Constitution states that everyone has the right to be free from all forms of violence. This right is broadly formulated, and therefore may be interpreted to include the right to be free from online violence. Further, and which will be discussed below, online harms may lead to offline violence which may similarly infringe on this right.
- **Equality:** The right to equality⁴³ could be infringed by an online harm, depending on the facts of the case. In *South African Human Rights Commission v Khumalo*, the Equality Court found that the Respondent's posts on social media amounted to hate speech and infringed upon the complainant's right to equality on the basis of race.⁴⁴ It has been seen that online harms increasingly have gendered dimensions and disproportionately affect women through threats of sexual violence, misogynistic disinformation campaigns and NCDII.⁴⁵ Therefore, it may be that a woman victim or survivor's right to equality is infringed by an online harm. It is necessary to keep the concept of intersectionality in mind when assessing whether a person's right to equality and non-discrimination has been infringed.
- **Children's rights:** Section 28 of the Constitution states that a child's best interests are of paramount importance in every matter concerning the child, and that they have the right to be protected from abuse or degradation. The best interests of the child is a dynamic concept, which requires assessment appropriate to each context.⁴⁶ Depending on the facts of the case, online harms may infringe upon this right. According to the United Nations Special Rapporteur on the Sale, Sexual Exploitation and Sexual Abuse of Children, digital environments pose significant risks of online sexual exploitation for children.⁴⁷ While online spaces are often not originally designed for children, they play a significant role in children's lives. Therefore, it must be ensured that in all actions regarding their use of digital environments, including in the adjudication of online harms, their best interests are the primary consideration.⁴⁸ As will be further discussed below, online harms may have particular effects on children due to their vulnerability and developing capacity. Therefore, their constitutional rights may be infringed differently by online harms than that of an adult.⁴⁹

42 *KS v AM* 2018 (1) SACR 240 (GJ) at para 63 (accessible [here](#)).

43 See above n37 at section 9.

44 *South African Human Rights Commission v Khumalo* 2019 (1) (SA) 289 (GJ) at paras 76 – 104 (accessible [here](#)).

45 African Commission on Human and Peoples' Rights, 'Resolution on the Protection of Women Against Digital Violence in Africa – ACHPR. Res. 522 (LXXII)' (2022) (accessible [here](#)).

46 General Comment No. 25 (2021) on Children's Rights in Relation to the Digital Environment CRC/C/GC/25 (2021) at para 12 (accessible [here](#)).

47 Report of the Special Rapporteur on the Sale, Sexual Exploitation and Sexual Abuse of Children A/79/122, submitted in accordance with Human Rights Council Resolution 52/26 (2024) (accessible [here](#)).

48 See above n46.

49 Media Monitoring Africa, 'Children's Rights Online: Towards a Digital Rights Charter' (2020) (accessible [here](#)).



- **Freedom of expression:** Section 16 of the Constitution states that everyone has the right to freedom of expression. Online harms discourage victims and survivors from freely using online platforms.⁵⁰ There are concerning trends globally that show that online harms are increasingly being used as a way to silence and intimidate journalists and human rights defenders, particularly women.⁵¹ Online harms may therefore have a secondary effect of preventing people from exercising their right to freely express themselves, without fear of retaliation, on social media platforms.
- **Access to information:** As set out above, online harms discourage participation on online platforms – it has a chilling effect on online discourse and the free flow of information.⁵² This may limit the general public’s right to access to information, as set out in section 32 of the Constitution, particularly if the victim or survivor is a journalist or human rights defender. Discouraging individual victims and survivors from using online platforms may also infringe on that individual’s ability to access information.

The psychological and social effects of online harms

Online harms may cause a variety of negative psychological and social effects. It is important for the judiciary to be aware of these effects of online harms on victims and survivors in order to grant appropriate relief and protection. The degree and exact nature of such effects is dependent on the harm in question, as well as the individual victim or survivor. The impacts of online harms can be felt immediately or experienced over time and be enduring.⁵³

- **Psychological effects:** Online harms may affect victims and survivors psychologically in a variety of ways.⁵⁴ Online harms may cause anxiety, fear and stress.⁵⁵ Online harms may contribute to, or cause, depression. It is vital to note that online harms may entrench these psychological effects to the extent of long-term mental health issues – such as Post Traumatic Stress Disorder or Chronic Anxiety. These psychological effects may be completely debilitating, resulting in victims and survivors losing their ability to enjoy their life.
- **Social consequences:** Online harms may lead to the loss of enjoyment of life due to the negative effects they may have on a victim or survivor’s social interactions. Online harms, due to the psychological effects that they may cause, as discussed above, may lead to victims and survivors reducing their participation in social life.⁵⁶ This may damage their inter-personal relationships. This can play out in social settings, as well as in terms of work relationships.

50 Centre for Human Rights, ‘Understanding Online Gender-based Violence in Southern Africa’ (2021) at pg 9 (accessible [here](#)).

51 The United Nations Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe Representative on Freedom of the Media, the Organization of American States Special Rapporteur on Freedom of Expression, and the African Commission on Human and Peoples’ Rights Special Rapporteur on Freedom of Expression and Access to Information, ‘Joint Declaration on Freedom of Expression and Gender Justice’ (2022) (accessible [here](#)).

52 Id.

53 See above n1 at pg 8.

54 Id.

55 See, for example, the effect of online harassment: South African Human Rights Commission, ‘Social Media Charter’ (2023) at pg 19 (accessible [here](#)).

56 M Storry and S Poppleton, ‘The Impact of Online Abuse: Hearing the Victim’s Voice’ *Victims Commissioner* (2022) at pg 7 and 24 (accessible [here](#)).



Depending on the type of work that they do, the damage caused to work relationships may negatively impact a victim or survivor's financial stability. When assessing social consequences of an online harm, the context of the victim or survivor's culture must be kept in mind. The unlawful and non-consensual exposure of certain sensitive information, for example, could lead to ostracization and even further harm in certain communities. Stigmatisation and social exclusion may be caused by certain online harms. Again, this may impact a victim or survivor's ability to enjoy their life and further infringe on their constitutional rights.

The physical and offline effects of online harms

Online harms may have effects on victims and survivors' physical health and can also manifest into offline harms.

- **Effects on physical health:** The psychological effects of online harms, which have been discussed above, may result in physical health consequences. For example, prolonged stress can weaken the immune system, which makes individuals more susceptible to illness.⁵⁷ Further, mental health issues (caused by an online harm) can manifest physically through a variety of symptoms such as headaches, fatigue and sleep-disturbances. The intensity and duration of these symptoms is specific to the individual involved. Depending on the type of online harm experienced, and the social consequences of such a harm, a victim or survivor may also turn to self-harming practices as a coping mechanism.⁵⁸
- **Offline manifestations:** Online harms may translate into offline harms and a further infringement of rights. For example, the non-consensual sharing of a person's private information, such as their address, may lead others who then have access to that information to stalk or physically harm the victim.⁵⁹ Further, online hate speech may be used by a person or group to incite others to engage in physical violence.⁶⁰ In *Economic Freedom Fighters and Others v Manuel*, the Constitutional Court held that the impugned tweet had the real life effect of damaging reputation, and accordingly ordered that damages were appropriate.⁶¹ It is clear that online harms may manifest into offline harms and, as such, the suitable degree of seriousness should be attached to online harms.

⁵⁷ F Dhabhar, 'Effects of stress on immune function: the good, the bad, and the beautiful' (2014) (accessible [here](#)).

⁵⁸ NSPCC, 'Online abuse' (2024) (accessible [here](#)).

⁵⁹ See above n55 at pg 16.

⁶⁰ There have been various examples of this around the world. In South Africa, campaigns such as Operation Dudula have used online platforms to incite xenophobic violence (see [here](#)).

⁶¹ *Economic Freedom Fighters and Others v Manuel* 2021 (3) SA 425 (SCA) at para 127 (accessible [here](#)).



The particular effect of online harms on children

Our courts have made it clear that children need support and assistance in order to enable their positive growth and development.⁶² Children are still developing and therefore need the appropriate care and protection to ensure that they reach their full potential.⁶³ The digital environment is becoming increasingly important across various aspects of children's lives and, according to the United Nations Committee on the Rights of the Child:

"The rights of every child must be respected, protected and fulfilled in the digital environment. Innovations in digital technologies affect children's lives and their rights in ways that are wide-ranging and interdependent, even where children do not themselves access the Internet".⁶⁴

Given this context, it is necessary to examine the particular effects of online harms on children. Online harms may have a devastating impact on the mental and physical wellbeing on children, due to children's evolving capacities and as they are generally more needful of protection, and less resourceful than adults.⁶⁵

- **Developmental vulnerability:** Children are in critical stages of their cognitive and emotional development. Their brains are still developing, which means that they are more susceptible to the psychological effects of online harms,⁶⁶ as has been set out above. Further, exposure to online harms may disrupt their cognitive and emotional development, leading to long-term psychological issues.⁶⁷ Children are generally not as equipped as adults to deal with negative experiences – they may lack the emotional resilience and coping mechanisms needed to handle online harms. This may result in heightened psychological and physical impacts.
- **Increased risk of exploitation:** Children may be more vulnerable to certain types of online harms, such as online predation and exploitation. Depending on the child in question, they may not fully understand the risk associated with sharing personal information or engaging with strangers.⁶⁸

These characteristics of children must be kept in mind when assessing the impact of online harms on them. It is important to note that children develop at different paces and have different life experiences, which may impact the way that online harms affect them.⁶⁹

62 *Teddy Bear Clinic for Abused Children and Another v Minister of Justice and Constitutional Development and Another* 2014 (2) SA 168 (CC) at para 1 ([accessible here](#)).

63 See above n14 at page 16.

64 See above n46 at para 4.

65 See above n49 at page 38.

66 J Bryce, S. Livingstone, J. Davidson, B. Hall and J. Smith, 'Evidence Review on Online Risks to Children' NSPCC (2023) at pg 57 ([accessible here](#)).

67 *Id* at pg 41.

68 S Ruddy, 'Online harms: protecting children and young people' NSPCC (2024) ([accessible here](#)).

69 See above n63 at pg. 17.



4.

An overview of the main social media platforms

Online harms occur mostly on social media platforms. Therefore, in order to fully understand online harms and how they manifest, it is necessary to understand how these platforms work. It is important to note that new social media platforms may be created and become popular at any time, and the nuances of each platform must be examined. The below section sets out the features and nuances of the platforms that are presently popular in South Africa – WhatsApp, X, Instagram, Facebook, and TikTok.

WhatsApp

WhatsApp is an application (“app”) through which people can send text messages, voice messages, photos, videos, documents and even locations to their contacts. Direct messages are end-to-end encrypted, meaning that only the sender and recipient can read the message.⁷⁰ In order for someone to be one of your contacts on WhatsApp, you must have the person’s phone number saved, and they must also have a WhatsApp account.⁷¹ In order to create an account, one needs to download the app and register with a valid phone number, as well as select a username. A user’s contacts will be able to see their phone number and username on their profile, depending on their privacy settings.⁷²

⁷⁰ WhatsApp Help Center, ‘About end-to-end encryption’ ([accessible here](#)).

⁷¹ WhatsApp Help Center, ‘How to find and invite contacts’ ([accessible here](#)).

⁷² WhatsApp Help Center, ‘How to change your privacy settings’ ([accessible here](#)).





Key terminology: read receipts and blocking

Read receipts is a feature on WhatsApp that indicates to a user's contacts whether they have read their messages or not. The symbol of one grey tick next to a message indicates that the message was successfully sent, two grey ticks indicates that the message was successfully delivered, and two blue ticks indicates that the recipient has read the message.⁷³ However, users may change their settings so that the ticks will always be grey, regardless of whether they have read the message or not.

On WhatsApp, users have the option of **blocking** someone. When a user blocks another, the blocked user can no longer send them messages, see their status, view their profile picture, or see their contact information.⁷⁴ The blocked user does not get notified that they have been blocked.⁷⁵

WhatsApp has several features that take it beyond an individual-to-individual messaging platform. These features must be examined to holistically understand how the app functions.

- **Disappearing messages:** Disappearing messages is an optional feature that users can turn on in their settings. Using this feature, users can set messages to disappear after 24 hours, 7 days, or 90 days.⁷⁶ After the message disappears, both the sender and receiver will no longer be able to see it. This may have adverse implications for evidence gathering in the case of an online harm, unless the victim or survivor screenshot the message before it disappeared.⁷⁷
- **Groups:** Users can create groups, with up to 1024 participants, on WhatsApp.⁷⁸ In groups, the user that created the group, or participants who have been designated as admin by the creator, can adjust the group settings so that everyone can share messages or media, or otherwise so that only the admin can do so.⁷⁹ Depending on the settings of the group, group admins may have additional controls such as adding or removing members and setting group descriptions.⁸⁰ The participants in a group are able to see all of the other participants' contact information, even if they do not have that person saved as a contact on their phone. However, this can be changed depending on the privacy settings of the group.
- **Communities:** On WhatsApp, users can create communities, which bring related groups together under one umbrella.⁸¹ A typical example is a suburb community, which consists of separate street groups. Every user that is in the suburb community will see what is shared in the community group, but only those in the separate street groups will see what is shared

⁷³ WhatsApp Help Center, 'How to check read receipts' (accessible [here](#)).

⁷⁴ WhatsApp Help Center, 'About reporting and blocking someone on WhatsApp' (accessible [here](#)).

⁷⁵ Id.

⁷⁶ WhatsApp, 'About disappearing messages' (accessible [here](#)).

⁷⁷ A screenshot is an image of what is shown on a computer or cell phone screen. It captures what was on the screen at the moment the screenshot was taken. See Cambridge Dictionary, 'Screenshot' (accessible [here](#)).

⁷⁸ WhatsApp Help Center, 'How to create and invite into a group' (accessible [here](#)).

⁷⁹ WhatsApp Help Center, 'How to manage group admins' (accessible [here](#)).

⁸⁰ Id.

⁸¹ WhatsApp Help Center, 'About communities' (accessible [here](#)).



there. The community settings will dictate whether participants are able to see each other's contact information.⁸² The purpose of communities is to aid in organising and managing multiple groups more efficiently.⁸³

- **Statuses:** Statuses are temporary posts that disappear after 24 hours.⁸⁴ Users can post text, photos or videos as a status. The user posting the status can choose which of their contacts are able to see their status through privacy settings.⁸⁵

If users experience abuse or inappropriate content while using WhatsApp, they can report individual contact or groups through an integrated reporting feature in the app.⁸⁶ Once reported, WhatsApp will receive the last 5 messages that the reported user or group sent, and then will take appropriate action based on their policies.⁸⁷

X (previously Twitter)

X is a social media platform where users can post and interact with messages. These messages are known colloquially as tweets, although since changing its name, X refers to them as posts. Ordinary users can post short messages, up to 280 characters, including text, images, videos and links.⁸⁸ In order to create an account, users need to provide a username, phone number and email address.⁸⁹ A user's username is known as their handle.⁹⁰ Further, users may adopt a profile picture, a short description about themselves (a bio), their location, and link to a website.⁹¹ Anyone who visits a user's profile is able to see their profile picture, name, bio and location.

82 WhatsApp Help Center, 'About general chat for communities' (accessible [here](#)).

83 See above n81.

84 WhatsApp, 'How to use status' (accessible [here](#)).

85 Id.

86 See above n74.

87 Id.

88 X, 'How to Post' (accessible [here](#)).

89 X, 'Signing up with X' (accessible [here](#)).

90 X, 'How to change your x Username' (accessible [here](#)).

91 See above n89.





Key terminology: timelines, private messages, and public or protected tweets

On X, users each have a **timeline** which displays the tweets of those that they follow.⁹² X's algorithms also identify tweets or accounts that it deems to be popular or relevant to the user in question and adds that content to users' timelines.⁹³

Users can send **private messages** or Direct messages ("DMs") to those that follow them or even those that do not, depending on their privacy settings.⁹⁴ Such private messages are only visible to the sender and receiver.

Users can adjust their settings so that their **tweets are public or protected**. Tweets are by default public, meaning that anyone can see them – even people who do not have an account. However, users may also set their tweets as protected, making them only visible to their followers.⁹⁵

X has a variety of features which can be used by users to interact with tweets. These are set out below.

- **Reposting and liking:** Users can repost tweets, which essentially means that the user re-shares an existing tweet.⁹⁶ Users can additionally add their own comment or media before reposting the tweet.⁹⁷ The user's followers will then be able to see the reposted tweet, as well as be able to go directly to the original tweet. Users can also repost their own tweets, which will bring the tweet to the user's followers' attention on their timelines again.⁹⁸ Users may also like tweets, which is an indication of admiration or endorsement. Viewers of the tweet will be able to see how many likes it has, as well as which users liked it.⁹⁹ The author of the tweet will be notified of all of the likes to their tweet.¹⁰⁰
- **Replies:** A reply is a response to another user's tweet¹⁰¹. Anyone who can view the original tweet will also be able to see all of the replies made on that tweet.¹⁰² Therefore, users will be able to see replies even if they do not follow the user who made the reply. The author of the original tweet is able to delete replies made on their tweet, and thereafter no one will be able to see the deleted reply.¹⁰³
- **Mentions:** Users can mention each other in tweets or replies. The effect of mentioning someone is that it creates a link to the mentioned user's profile, and notifies them about the tweet or reply in question.¹⁰⁴ A user who has been mentioned in a tweet can untag themselves from unwanted tweets or replies.¹⁰⁵

92 X, 'About your for you timeline on X' (accessible [here](#)).

93 Id.

94 X, 'About Direct Messages' (accessible [here](#)).

95 X, 'About public and protected Posts' (accessible [here](#)).

96 X, 'How to Repost' (accessible [here](#)).

97 Id.

98 Id.

99 X, 'How to like a post' (accessible [here](#)).

100 Id.

101 X, 'About replies and mentions' (accessible [here](#)).

102 Id.

103 Id.

104 Id.

105 Id.



- **Threads:** A thread on X is a series of connected tweets from one user.¹⁰⁶ Users can provide additional context, an update to a tweet, or an extended point, through connecting multiple tweets together through a thread.¹⁰⁷ Users will be able to see all of tweets within a thread, as well as all of the likes, reposts, replies, and mentions from each individual tweet.¹⁰⁸

On X, users are able to report tweets, replies, or private messages that are in violation of the platform's rules or terms of service.¹⁰⁹ For example, users can report tweets that contain abusive content. X will then review the reported content and, if it deems that the content is in violation of its policies, they will take action.¹¹⁰ Action in this sense may result in anything from a warning to the perpetrating user, to a suspension of their account.

Instagram

Instagram is a social media platform that enables users to share and interact with photos and videos. In order to create an account, users must be above 13 years of age, provide a valid phone number, email address, and a name (which can be real or a pseudonym).¹¹¹ Users may also elect to upload a profile picture, description of themselves, and a link to a personal or professional website.¹¹² Anyone who visits a user's profile is able to see their profile picture, name, bio, and website.¹¹³



Key terminology: followers, feeds, and public or private accounts

The concept behind Instagram is that users **follow** each other. If you follow a user, you are known as their follower.¹¹⁴

Once a user follows another, their posts will be displayed their **feed**.¹¹⁵ A user's feed is where all the posts from the people that they follow, as well as advertisements that Instagram's algorithm thinks that you would be interested in, are displayed. A user's feed is the first thing that they see when they go into the app.

Users can elect whether to make their **account private or public**. If an account is public, it means that any user who visits their profile can see what they have posted and may follow that user.¹¹⁶ If a user's profile is private, it means that others will only be able to see their posts and follow them if that user has given their permission.¹¹⁷

¹⁰⁶ X, 'How to create a thread on X' (accessible [here](#)).

¹⁰⁷ Id.

¹⁰⁸ Id.

¹⁰⁹ X, 'Report abusive behavior' (accessible [here](#)).

¹¹⁰ Id.

¹¹¹ Instagram Help Center, 'Your Profile' (accessible [here](#)).

¹¹² Id.

¹¹³ Id.

¹¹⁴ Instagram Help Center, 'Syncing Contacts and Finding People to Follow' (accessible [here](#)).

¹¹⁵ Instagram Help Center, 'How Instagram Feed Works' (accessible [here](#)).

¹¹⁶ Instagram Help Center, 'Differences between public and private accounts on Instagram' (accessible [here](#)).

¹¹⁷ Id.



Instagram has developed to allow users to post a variety of content and to interact with content in a variety of ways. This is set out below.

- **Photos, videos, or reels:** On Instagram, users can post a photo, video, or a series of photos and videos, either with or without a caption. A series of photos or videos in a singular post is known as a carousel.¹¹⁸ Further, users may also record, edit and then post short videos known as reels. Users can use the reels function to merge existing videos clips, record new video clips, and add music and other effects to the clips.¹¹⁹ Reels may similarly be post with or without captions.
- **Stories:** Users can share photos and videos that will disappear after 24 hours. These temporary posts are known as stories.¹²⁰ However, users can elect to add a story to their highlights – which means that the story will then be visible on the user’s profile even after the 24 hour period has lapsed.¹²¹ Users can also elect to only share a story with their close friends. Users can choose which of their followers is in their close friends list, and only those users will be able to see the story in question.¹²²
- **Live:** Users can live broadcast to their followers. Everyone who follows the user will be able to see in real time whatever the broadcaster’s camera is filming at that moment.¹²³ Once a live broadcast has ended, only the broadcaster will be able to share a replay of it or access it in their archive.¹²⁴
- **Comments:** If a user has a public profile, anyone may comment on their posts. If the user has a private profile, only those following them may comment on their posts.¹²⁵ Any user that is able to view the post, will also be able to see all of the comments made and also is able to reply to any comment that has been made. The author of the post is able to delete comments and replies thereto. Users may also elect to turn off comments.¹²⁶ Thereafter others will not be able to comment on their posts.
- **Tags and mentions:** Users can “tag”¹²⁷ other users in the photos, carousels, reels, or stories that they post.¹²⁸ If a user has a private profile, they will only be able to tag their followers. Anyone who is able to see the post will be able to see who has been tagged in said post. The user who has been tagged will get a notification of the same and the post will be added to the tagged user’s profile.¹²⁹ However, if the tagged user has a private profile, only their approved followers will be able to see posts that they have been tagged in.¹³⁰ If a user is tagged in a story,

118 Instagram Help Center, ‘Sharing Photos and Videos’ (accessible [here](#)).

119 Instagram Help Center, ‘Reels’ (accessible [here](#)).

120 Instagram Help Center, ‘Stories’ (accessible [here](#)).

121 Id.

122 Instagram Help Center, ‘Create a Close Friends list on Instagram’ (accessible [here](#)).

123 Instagram Help Center, ‘Live’ (accessible [here](#)).

124 Id.

125 See above n 116.

126 Instagram Help Center, ‘Turn comments on or off for Instagram posts’ (accessible [here](#)).

127 To tag something means to use a word/s or phrase to categorize content and make it easier for users to find and engage with specific topics.

128 Instagram Help Center, ‘Tagging and Mentions’ (accessible [here](#)).

129 Id.

130 Id.



they will be able to re-post the story. Users can also mention other users in a comment. The effect of mentioning another user in a comment is that that user will then get a notification of the same, drawing their attention to the post in question.¹³¹ If the user who posted the post has a private account, only those who follow them will be able to see the original post, even if they have been mentioned in a comment. Instagram has a range of settings to manage who can tag or mention a user in a post. For example, a user could choose to only allow tags or mentions from people that they follow.¹³²

- **Direct messages:** Instagram allows users to send direct messages to other users, or to a group of users.¹³³ These messages can include text, photos, and videos. Users can also send profiles or posts to other users.¹³⁴ Users may adjust their settings so that only users that they follow can send them messages directly. Messages that are sent from users that they do not follow will be added to a requests folder – the user will have to accept the message before it appears in the user’s chats.¹³⁵
- **Explore:** Besides seeing posts, stories, and live broadcasts from users that one follows, Instagram also offers a feature known as the explore page. The explore page is curated by Instagram’s algorithms, pulling together posts from users that the user does not follow, based on users that they do follow and posts that they interact with.¹³⁶ For example, if a user primarily follows users who post travel content, their explore page will most likely comprise of similar travel content from other users. The explore page is also where users can search for profiles, or specific types of content.¹³⁷

If users think that a post, message, or story does not follow Instagram’s community guidelines or terms of use, they can report it using the app’s embedded reporting features.¹³⁸ If Instagram finds that the content does violate their community guidelines or terms of use, it will hide or remove that content from the app.¹³⁹ Should Instagram find that there was not a violation, the user will have an opportunity to request another review.¹⁴⁰ Users can also report a profile if that profile posts content that it should not, is impersonating someone else, or is a child below the age of 13.¹⁴¹

131 Id.

132 Instagram Help Center, ‘Who can mention or tag you on Instagram’ (accessible [here](#)).

133 Instagram Help Center, ‘Send, View and Manage Messages’ (accessible [here](#)).

134 Id.

135 Id.

136 Instagram Help Center, ‘Exploring Photos and Videos’ (accessible [here](#)).

137 Id.

138 Instagram Help Center, ‘Report a post or profile on Instagram’ (accessible [here](#)).

139 Instagram Help Center, ‘What to do if Instagram didn’t remove content you reported’ (accessible [here](#)).

140 Id.

141 See above n137.



Facebook

Facebook is a social media platform that allows users to share updates, photos, and videos with their friends, and join communities of interest. Instagram and Facebook are owned by the same company, Meta, and as a result share many of the same features. In order to create an account, a user must provide their first and last name, phone or email address for verification, a password, their gender, and a date of birth.¹⁴² Users must be at least 13 years old.¹⁴³ Users may also provide a profile picture, short description of themselves, and a cover photo.¹⁴⁴ Visitors to a user's profile will be able to see their name, cover and profile pictures, and description. However, depending on the user's privacy settings, visitors may also be able to see additional information such as their contact details and birth date.¹⁴⁵



Key terminology: friends, feeds, and timelines

On Facebook, a user's connections are known as their **friends**. Friending someone refers to adding someone as a connection on Facebook. In order for users to become friends on Facebook, one of them must add the other, and that user must accept the friend request. Once added as a friend, users will be able to see each other's activity on Facebook.¹⁴⁶

A user's **feed** is where all of the content that their friends, and the pages and groups that they follow, post will appear.¹⁴⁷ A user's feed is the landing page of Facebook, and is the first thing that they will see when opening the app.

A user's **timeline** is a section of their profile where one can see all of their posts and activities in chronological order. Posts from user's friends that they are tagged in will also be displayed on their timeline.

Facebook's main features work very similarly to that of Instagram. Users can post photos,¹⁴⁸ videos, reels,¹⁴⁹ stories,¹⁵⁰ and live broadcast.¹⁵¹ Other users can then share, comment on, or like this content. Differently to Instagram, Facebook users may also post text without an accompanying photo or video. The visibility of Facebook posts is dependent on the author's privacy settings – users may dictate that only their friends can see their posts, or otherwise that their posts are visible to anyone who views their profile. While many of the features of Facebook are similar to Instagram, there are some different features, that are set out below.

¹⁴² Facebook Help Center, 'Create a Facebook Account' (accessible [here](#)).

¹⁴³ Id.

¹⁴⁴ Id.

¹⁴⁵ Facebook Help Center, 'Your Privacy' (accessible [here](#)).

¹⁴⁶ Facebook Help Center, 'Friending' (accessible [here](#)).

¹⁴⁷ Facebook Help Center, 'How Feed Works' (accessible [here](#)).

¹⁴⁸ Facebook Help Center, 'Photos' (accessible [here](#)).

¹⁴⁹ Facebook Help Center, 'Reels' (accessible [here](#)).

¹⁵⁰ Facebook Help Center, 'Stories' (accessible [here](#)).

¹⁵¹ Facebook Help Center, 'Go live on Facebook' (accessible [here](#)).



- **Tagging:** Similarly to Instagram, users may tag others in their posts or in comments.¹⁵² However, when users tag someone on Facebook, that post will then be visible to both the user tagged and their friends¹⁵³. Users can opt to hide the post in question from the friends of the user tagged.¹⁵⁴ Users can adjust their settings so that they can review tags before they are allowed to appear on their timeline, feed, or elsewhere on Facebook.¹⁵⁵
- **Facebook Dating:** Facebook Dating is essentially a dating app under the Facebook umbrella. At the moment, Facebook dating is not available in South Africa.¹⁵⁶ Through users' Facebook profiles, they can create a Dating profile.¹⁵⁷ Users with a Dating profile may then like another user that has also has a Dating profile and, once that user has liked them back, they can direct message each other through their Dating profiles.¹⁵⁸ While a user's Dating profile is created through their Facebook account, it is kept separate – a user's Facebook friends and anyone outside of Facebook Dating cannot see the user's Dating profile.¹⁵⁹ Users are able to delete their Dating profiles without deleting their Facebook account, but deleting their Facebook account would delete their Dating profile.¹⁶⁰
- **Events:** All users can create online or in person events on Facebook.¹⁶¹ Users can specify the relevant details of the event, such as the date, time, and location. The event can either be set as private or public. If it is private, only those who the event creator invites will be able to see the event itself as well as the information about the event. If it is public, all users will be able to see it and the information about the event.¹⁶² Users are also able to search for events that are happening near to them, specific types of events, and events that their friends are going to if those events are set as public.¹⁶³
- **Pages:** A Facebook Page is a public profile that is created for businesses, organisations, public figures, brands, or anyone seeking to promote themselves through Facebook.¹⁶⁴ Users can like or follow a page to receive updates from that page, which will appear on their feed. Pages can post or create events in the same way that individual users can, which will then be visible to those that follow the page.¹⁶⁵
- **Groups:** On Facebook, users can create groups centring on different themes or interests.¹⁶⁶ The user who created the group is known as the Group Admin, who then controls the functioning of the group.¹⁶⁷ The Group Admin may also delegate this control to other members of the group. Depending on the group's settings, it may be an open or closed group. If it is an open group,

¹⁵² Facebook Help Center, 'Tagging' (accessible [here](#)).

¹⁵³ Id.

¹⁵⁴ Id.

¹⁵⁵ Id.

¹⁵⁶ Facebook Help Center, 'Facebook Dating' (accessible [here](#)).

¹⁵⁷ Id.

¹⁵⁸ Id.

¹⁵⁹ Id.

¹⁶⁰ Id.

¹⁶¹ Facebook Help Center, 'Events' (accessible [here](#)).

¹⁶² Facebook Help Center, 'Control who sees or joins your Facebook event' (accessible [here](#)).

¹⁶³ Facebook Help Center, 'Find Facebook events you might be interested in' (accessible [here](#)).

¹⁶⁴ Facebook Help Center, 'Pages' (accessible [here](#)).

¹⁶⁵ Id.

¹⁶⁶ Facebook Help Center, 'Groups' (accessible [here](#)).

¹⁶⁷ Facebook Help Center, 'Join a Facebook group' (accessible [here](#)).



any user may join. If it is a closed group, the Group Admin will need to approve all requests to join the group.¹⁶⁸ A group may post or create events in the same way as an individual user, and all those who are part of the group will be able to see its posts on their feeds.

- **Marketplace:** Facebook Marketplace is a platform embedded within Facebook where users can buy and sell items.¹⁶⁹ Users interested in buying items on Marketplace can search per category or for something specific.¹⁷⁰ Users can filter the results through price range or by location. In selecting an item, a user will be able to directly message the seller as well as view the seller's profile.¹⁷¹ If a user wishes to sell something on Marketplace, they can list the item.¹⁷² They will be required to give a description of the item, the price it is being sold for, and an approximate location of the item.¹⁷³ All listings are public, and can be seen by anyone on Marketplace.¹⁷⁴

On Facebook, users may make use of the app's embedded reporting function to report abusive content.¹⁷⁵ Once content is reported, Facebook reviews it and removes it if it goes against its community standards.¹⁷⁶

TikTok

TikTok is a social media app that enables users to create and share short form videos. In order to create an account, users need to provide a username, and valid email addresses and phone number,¹⁷⁷ and must be at least 13 years old.¹⁷⁸ Users may additionally select a profile picture and nickname. Visitors to users' profiles will be able to see their username, profile picture, and nickname.¹⁷⁹ Users are able to choose between a private or public account. If a user chooses a private account, they will have control over who is able to follow them, watch their content, and who can view who the user follows.¹⁸⁰ On the other hand, a public account will be visible to anyone on or off TikTok, unless the user adjusts their privacy settings.¹⁸¹

168 Id.

169 Facebook Help Center, 'Marketplace' (accessible [here](#)).

170 Facebook Help Center, 'Find items to buy on Facebook Marketplace' (accessible [here](#)).

171 Id.

172 Facebook Help Center, 'Selling on Marketplace' (accessible [here](#)).

173 Id.

174 Id.

175 Facebook Help Center, 'Report Content on Facebook' (accessible [here](#)).

176 Facebook Help Center, 'Reporting Abuse' (accessible [here](#)).

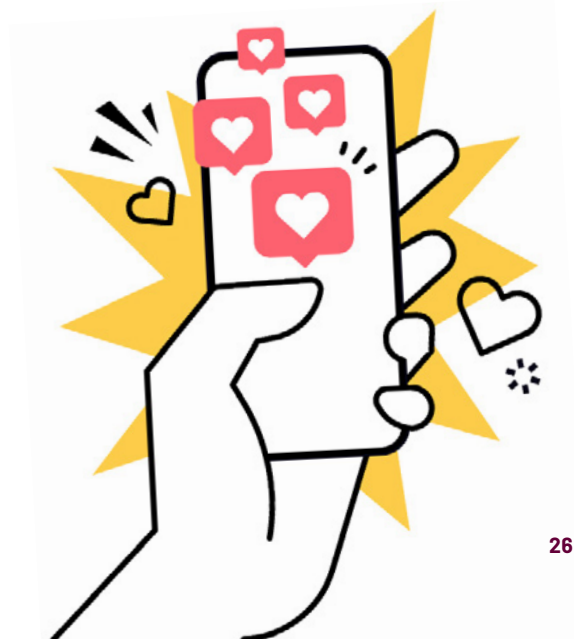
177 TikTok, 'Creating an account' (accessible [here](#)).

178 TikTok, 'Youth Safety and Well-Being' (accessible [here](#)).

179 TikTok, 'Choosing between a private or public account' (accessible [here](#)).

180 Id.

181 Id.





Key terminology: hearts and for you

Hearts are how users show their admiration for a post. It is the equivalent of a like on Facebook. Users will get notified when another user hearts their content, and other users are able to see who hearted a post.¹⁸²

Users **for you** pages are personalised feeds of content, based on the user's interests and engagement. It is the first thing that users see when they open the app, and is comparable to the explore page on Instagram.¹⁸³

Users can elect to either create videos through the app, or otherwise upload existing videos from their devices.¹⁸⁴ Users may then post the video or keep it in drafts to post at a later date.¹⁸⁵ Users may elect to add a caption to the video, or not.¹⁸⁶ Once posted, the visibility of the video is dependent on whether the user has elected to have a public or private profile, and their privacy settings, as set out above. As with other social media platforms, there are particular features of the app that require explanation.

- **Tagging and comments:** Users can personalise their privacy settings to specify who can comment on their videos.¹⁸⁷ For example, users may elect to only allow users who they follow and who follow them back to comment on their videos. Comments are then visible to all users who are able to view the original post. Users may also elect to not allow comments for specific videos.¹⁸⁸ Users can similarly control who may tag them in posts, stories, and comments.¹⁸⁹ Once tagged, all those who can view the original post, story, or comment will be directed to the tagged user's profile.¹⁹⁰ Whether they can then view the content of the tagged user's profile is dependent on the tagged user's privacy settings.¹⁹¹
- **TikTok Stories:** TikTok Stories allow users to share photos and videos with their followers for 24 hours (or to anyone, if the user's profile is public).¹⁹² Users who are able to view these temporary posts are then able to react and send replies to the story, which only the author of the story will be able to view.¹⁹³
- **Direct messages:** Direct messaging allows users to privately share text, photos, or videos with each other.¹⁹⁴ Users can send each other public posts of other users, as well as novel text, images, or videos. Users can control who is able to directly message them through their privacy settings.¹⁹⁵

182 TikTok, 'TikTok Glossary Part 1' (accessible [here](#)).

183 TikTok, 'For You' (accessible [here](#)).

184 TikTok, 'Making a post' (accessible [here](#)).

185 Id.

186 Id.

187 TikTok, 'Comment' (accessible [here](#)).

188 Id.

189 TikTok, 'Mentions on TikTok' (accessible [here](#)).

190 Id.

191 Id.

192 TikTok, 'TikTok Stories' (accessible [here](#)).

193 Id.

194 TikTok, 'Direct Messages' (accessible [here](#)).

195 Id.



- **Duets:** Users with public accounts can Duet with other users that also have public accounts.¹⁹⁶ A Duet contains two videos that display in a split screen, that play at the same time.¹⁹⁷ Users can control in their privacy settings who is permitted to Duet their videos.¹⁹⁸ For example, users can choose to only allow users who they follow and who follow them back to Duet with their videos. If a user Duets with another's video, the Duet will only appear on the profile of the person who recorded the Duet.¹⁹⁹
- **Stitch:** Stitch allows users to combine existing videos on TikTok from public users with one that they create.²⁰⁰ A Stitch is different from a Duet in that it displays in a sequence. First the original video will play, followed by the stitched content. In a Duet, the two videos display side by side. Similarly to Duets, users can control through their privacy settings who can Stich their videos.²⁰¹ The Stitched video will only appear on the profile of the user who recorded the Stitch.²⁰²

TikTok users may report content through the app's embedded reporting mechanism.²⁰³ If the content violates TikTok's policies, they will remove the content after they have reviewed it.²⁰⁴ Depending on the legality of the content, TikTok may elect to remove it globally or only in the country where it is illegal.²⁰⁵

Community standards, reporting mechanisms, transparency of processes, and appeal mechanisms all differ significantly from platform to platform. Each platform has its own set of guidelines and procedures for handling potentially harmful content, which can lead to a lack of consistency in how issues such as hate speech and incitement are addressed. For example, some platforms may have more stringent policies and quicker response times, while others might be less transparent about their processes and slower to act on user reports.

In South Africa, some civil society organisations are developing tools to ensure consistency in the way that online harms are reported and thereafter processed. For example, users may report problematic content to Real411. Real411 is an independent platform where South African users can lodge complaints about digital misinformation, disinformation, and other online harms.²⁰⁶ This platform allows users to report content they believe violates community standards or legal provisions related to hate speech and incitement, providing an additional layer of oversight and accountability beyond the internal mechanisms of individual platforms. By utilising Real411, South Africans can contribute to a more coordinated and effective approach to combating online harms within the national context.

¹⁹⁶ TikTok, 'Duets' (accessible [here](#)).

¹⁹⁷ Id.

¹⁹⁸ Id.

¹⁹⁹ Id.

²⁰⁰ TikTok, 'Stitch' (accessible [here](#)).

²⁰¹ Id.

²⁰² Id.

²⁰³ TikTok, 'Reporting Illegal Content' (accessible [here](#)).

²⁰⁴ Id.

²⁰⁵ Id.

²⁰⁶ Real 411 (accessible [here](#)).



5.

Existing legislation

As we aim to provide comprehensive guidance to judges on handling cases related to online harms, it is crucial to understand the existing legal framework in South Africa. This section will delineate the relevant legislation that addresses various aspects of online harms, ensuring that judicial officers are well-equipped to navigate this complex and evolving area of law. The legislation to be discussed includes the Cybercrimes Act, which addresses modern cyber threats; the Electronic Communications and Transactions Act (“ECTA”), which regulates electronic communications and transactions; the Protection from Harassment Act (“PHA”), which provides measures against harassment, including online harassment; the Protection of Personal Information Act (“POPIA”), which safeguards personal data; the Promotion of Equality and Prevention of Unfair Discrimination Act (“PEPUDA”), which promotes equality and prevents discrimination; the Prevention and Combating of Hate Crimes and Hate Speech Act, which addresses hate crimes and hate speech; and the Film and Publication Act. This comprehensive overview will assist in understanding the legislative landscape, applying the law effectively, and ensuring justice in cases involving online harms.

The Cybercrimes Act and Electronic Communications and Transactions Act

The Cybercrimes Act 19 of 2020²⁰⁷ significantly extends the offences outlined in the Electronic Communications and Transactions Act 25 of 2002²⁰⁸ encompassing a broader range of activities related to the illegal use of computer systems. The Cybercrimes Act is organised into nine chapters. Chapter 2 details various forms of cybercrimes, malicious communications, sentencing, and protective orders for victims of harmful communications, while Chapter 4 outlines the powers of law enforcement to investigate, search, access, or seize relevant data.

²⁰⁷ Cybercrimes Act 19 of 2020 ([accessible here](#)).

²⁰⁸ Electronic Communications and Transactions Act 25 of 2002.



Chapter 8 sets out the obligations of electronic communications service providers and financial institutions to report cybercrime offences and preserve relevant information for investigations.

The preamble of the Cybercrimes Act states that its aim is to create offences related to cybercrime and prescribe penalties for such crimes. The Act amends eleven critical pieces of legislation, including the Criminal Procedure Act, the South African Police Services Act, the Films and Publications Act, and the Financial Intelligence Centre Act, among others. This extensive list of amendments underscores the Act's comprehensive approach to updating South Africa's legal framework to address the complexities of modern cybercrimes.

The Cybercrimes Act categorises cybercrimes into different generations. First-generation crimes are traditional crimes assisted by technology, such as theft or fraud, where computers are used incidentally.²⁰⁹ Second-generation crimes are facilitated by computer networks, including the dissemination of child abuse images and fraud. These crimes can occur without computers but are significantly amplified by their use.²¹⁰ Third-generation crimes are true cybercrimes where computers or networks are the targets, such as hacking, denial of service attacks, and the distribution of malicious software.²¹¹

The Cybercrimes Act criminalises various activities. Section 2 covers unlawful access to computer systems or data storage mediums. Section 3 deals with the unlawful interception of data, and Section 4 addresses the unlawful use of software or hardware tools. Section 5 pertains to the unlawful interference with data or computer programs, while Section 6 focuses on the unlawful interference with computer data storage mediums or systems. Section 7 criminalises the unlawful acquisition, possession, provision, receipt, or use of passwords, access codes, or similar data. Other sections address cyber fraud, cyber forgery and uttering, cyber extortion, and theft of incorporeal property.

Section 2 of the Cybercrimes Act specifies two types of offences related to unlawful access. Firstly, it criminalises intentional actions involving a computer system or data storage medium that enable the commission of an offence. Secondly, it prohibits unauthorised access to a computer system or data storage medium. Unlawful access, commonly known as hacking, is defined as gaining unauthorised access to a computer system, programs, or data. Different branches of the South African government, such as the South African Police Services and the Government Communication and Information System, have been targeted by hacking attacks, highlighting the importance of robust cybercrime legislation. The Act criminalises the use of hacking tools like eBlaster for illicit purposes, such as siphoning funds from organisational bank accounts.

The Cybercrimes Act and ECTA have led to several significant prosecutions that illustrate its application and the judiciary's interpretation of its provisions.²¹²

²⁰⁹ See above n208 at Chapter 2.

²¹⁰ *Id.*

²¹¹ *Id.*

²¹² See *Okundu v S* (CA&R117/16) (2016) ZACGHC 131; *R v Douvenga* (unreported); *Mgoqi v S* (CA&R 46/2017) (2020) 33; *Myeni v S* (unreported); *Salzmann v S* (2019) 1 All SA 361 (SCA); 2020 (2) SACR 200 (SCA).





Relevant case law

- *Okundu v S* involved the illegal accessing of data, resulting in prosecution under section 86 of ECTA. The case demonstrated the judiciary's commitment to upholding the integrity of digital data and penalising unauthorised access.
- In *R v Douvenga* the accused was found guilty under section 86(1) of the ECTA for unlawfully accessing confidential data and emailing it to her fiancé. This case highlights how personal relationships can sometimes play a role in cybercrime, with individuals using their access to sensitive information for personal gain or to harm others.
- In *Mgoqi v S*, the court dealt with unauthorised access, reinforcing the legal stance against hacking and the importance of protecting digital infrastructures.
- *Myeni v S* involved data interference, constituting a denial of service attack under section 86(5) of ECTA. This case underscored the severe implications of denial-of-service attacks on public and private entities, emphasising the need for stringent penalties to deter such activities.
- In *Salzmann v S* the court found the offence under section 86(5) of ECTA to be serious, leading to a significant judgment that highlighted the judiciary's approach to tackling sophisticated cybercrimes.

The Cybercrimes Act and ECTA represent crucial steps in addressing the modern challenges posed by cybercrime. By expanding the scope of criminalised activities and establishing comprehensive enforcement mechanisms, South Africa aims to enhance its ability to combat cyber threats and protect the integrity of its digital infrastructure.

The Protection of Personal Information Act

The Protection of Personal Information Act 4 of 2013²¹³ serves as South Africa's primary legislation governing data protection. It sets guidelines for how both public and private entities may handle personal information, establishing eight core principles for lawful data processing. These principles emphasise accountability, transparency, and clearly defined purposes for information collection. Organisations are mandated to secure informed consent from individuals prior to gathering, utilising, or sharing their personal data, and they must implement adequate security measures to prevent unauthorised access or breaches.

POPIA empowers individuals with specific rights, such as the ability to access their own personal data and to object to its processing. These rights are fundamental to ensuring that individuals maintain control over their personal information. An independent body known as the Information Regulator is tasked with overseeing compliance with POPIA. This regulator provides guidance on the law's application, helping to clarify the responsibilities of organisations under this framework. The applicability of POPIA to social media platforms presents some ambiguities.

213 Protection of Personal Information Act 4 of 2013 (accessible [here](#)).



Two critical provisions warrant consideration. First, Section 3 of POPIA outlines when the Act applies, specifically addressing personal data processing by entities domiciled in South Africa or by those operating within the country. This raises questions for foreign social media companies, such as whether their operations in South Africa constitute data processing under POPIA. For POPIA to be applicable, these companies must actively process personal information within the country, which includes activities beyond merely forwarding data.

Second, POPIA's provisions may not extend to information that individuals have intentionally made public. However, there is a lack of judicial guidance on what constitutes "public" information within the context of the Act. While some jurisdictions treat public data as exempt from protection, others maintain that privacy principles still apply even to information shared on social media. Given the complex nature of social media operations, where many companies do not clearly disclose their data processing methods, there is an urgent need for clarity regarding POPIA's application to these platforms. For instance, while a company like Meta has a presence in South Africa, it remains uncertain how it processes personal information in the country.

There is a need for a clear understanding of both the rights of individuals regarding their personal information and the responsibilities of organisations, particularly those operating online. The ambiguity surrounding data processing can lead to potential violations of individuals' rights and the effective enforcement of data protection laws. Judges should be aware that the evolving nature of online interactions may necessitate a more adaptive interpretation of data protection laws, considering both the rights of individuals and the practicalities of enforcing these rights in a digital landscape.

The Protection from Harassment Act

The Protection from Harassment Act 17 of 2011²¹⁴ addresses various forms of harassment, including online harassment or cyberbullying. It was introduced to provide victims of harassment with a legal tool to protect themselves and to seek relief, particularly through obtaining protection orders. Although not explicitly designed for online harms, the PHA has provisions that apply to online harassment. The Act defines harassment broadly, including both physical and electronic forms of abuse. The law covers behaviours that cause harm, either directly or indirectly, through communication or by sending messages that induce fear, harm, or distress.

This means that any communication via digital platforms, such as social media, text messages, or emails, could be considered harassment if it meets the legal threshold. Victims and survivors of online harassment can apply for a protection order against the perpetrator. The Act includes "electronic communications" in its scope, which covers any written, visual, or recorded communication sent over electronic platforms. This includes text messages, social media posts, emails, and other online interactions that cause mental, emotional, or psychological harm. Additionally, the Act provides some mechanisms to address harassment by anonymous individuals, which is common in online spaces. The victim and survivor can request the court to

214 Protection from Harassment Act 17 of 2011 (accessible [here](#)).



issue a directive to an electronic communications service provider to disclose the identity or address of the person responsible for sending harassing electronic communications.

The Act provides protection for any individual, not just within intimate or domestic relationships. This is particularly relevant for online spaces, where harassment can occur between strangers or acquaintances, such as on social media or other digital platforms.

While the Act offers a framework to deal with online harassment, it has limitations:

- **Group harassment:** The Act is individualistic, meaning it does not cater to instances where harassment is perpetrated by a group of people acting together on platforms like social media.
- **Organisational harassment:** It does not protect organisations or groups targeted by harassment. Each individual would have to apply for a protection order, which can be costly and time-consuming.

In conclusion, the Act does offer mechanisms to address online harms in South Africa, but certain gaps, such as collective or group harassment and organisational targeting, still exist. There may be a need for future reforms to better address the evolving nature of online harassment.

The Promotion of Equality and Prevention of Unfair Discrimination Act and the Prevention and Combating of Hate Crimes and Hate Speech Act

The Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000²¹⁵ and the Prevention and Combating of Hate Crimes and Hate Speech Act 16 of 2023²¹⁶ both prohibit hate speech and incitement, including that which occurs online.

Section 10 of PEPUDA states that no person may publish, propagate, advocate or communicate works dated on a prohibited ground, against any person, that could reasonably be construed to demonstrate a clear intention to be hurtful; be harmful or incite harm; or promote or propagate hatred. The Constitutional Court in *Qwelane v South African Human Rights Commission and Another* narrowed the scope of this definition of hate speech.²¹⁷ Therefore, hate speech in terms of PEPUDA is now only that which could reasonably be construed to demonstrate a clear intention to be harmful or incite harm, or promote or propagate hatred. In terms of the Constitutional Court order, Parliament should amend PEPUDA to reflect the above.

Recently, there have been several cases which grapple with the concept of privacy and the application of PEPUDA in the context of hate speech which occurs through messaging platforms. For example, in *South African Human Rights Commission and Another v Lagardien* the Equality Court is tasked with deciding whether messages which were sent to a WhatsApp group constitute hate speech.²¹⁸

215 Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000 (accessible [here](#)).

216 Prevention and Combating of Hate Crimes and Hate Speech Act 16 of 2023 (accessible [here](#)).

217 *Qwelane v South African Human Rights Commission and Another* [2021] ZACC 22; 2021(6)SA 579 (CC) (accessible [here](#)).

218 *Power and Associates Inc., 'South African Human Rights Commission v Lagardien'* (2024) (accessible [here](#)).



The Prevention and Combating of Hate Speech Act, although not yet operational, criminalises hate speech. Here, hate speech is defined as it was by the court in *Qwelane*. In section 4(b) of the Act, hate speech by way of electronic communication is specifically identified as a form of prohibited conduct. Electronic communication is widely defined to mean a communication by means of data messages.²¹⁹ Therefore, hate speech and incitement that is communicated through social media platforms is criminalised.

PEPUDA and the Hate Speech Act substantially prohibit the same conduct. Therefore, one could face a civil and criminal sanction for the same instance of hate speech or incitement. How these Acts interact with each other once the Hate Speech Act becomes operational must be monitored.

The Film and Publication Act

The Film and Publication Act 11 of 2019²²⁰ empowers the Film and Publication Board to regulate speech by determining what constitutes incitement of imminent violence, propaganda for war, and advocacy of hatred. One of the primary objectives of the Act is to address the distribution of online content to ensure it aligns with established standards that prevent the dissemination of harmful material. The Act allows the Film and Publication Board (“FPB”) to regulate and monitor online platforms to protect the public from exposure to harmful or illegal content.

Specifically, the Act includes provisions that address the incitement of imminent violence, propaganda for war, and advocacy of hatred based on race, ethnicity, gender, religion, or nationality. According to Section 18F of the Act, the FPB is empowered to classify content that promotes or incites harm, violence, or hatred, potentially leading to its restriction or removal from online platforms. This provision ensures that content inciting such negative behaviours is subject to strict review and control. Additionally, the Act imposes compliance requirements on online distributors, including social media platforms. These platforms are required to register with the FPB and ensure their content adheres to the guidelines and classifications set by the Board. Non-compliance with these regulations can result in penalties, including fines or other legal actions, as stipulated in Section 24C. This measure aims to hold online distributors accountable for the content they host and distribute, ensuring a safer online environment.

The Act also emphasises public protection by preventing the spread of harmful content that could incite violence or hatred. Section 18H empowers the FPB to act against the dissemination of such content, ensuring that the rights to dignity, equality, and safety are upheld, particularly for vulnerable groups. By regulating online content, the FPB seeks to mitigate the social harms associated with hate speech and incitement. Furthermore, the FPB is tasked with promoting awareness and education regarding the responsible consumption and distribution of online content. This educational role is crucial in helping the public understand the impact of harmful content and the importance of adhering to content standards. Through these efforts, the Act aims to foster a more informed and responsible approach to online content distribution, reducing the prevalence of hate speech and incitement in digital spaces.

²¹⁹ See above n218 at section 1.

²²⁰ Film and Publication Act 11 of 2019 (accessible [here](#)).



6. Practical considerations

This section highlights the practical considerations to keep in mind when adjudicating online harms. These considerations are important to contemplate as each case will have a unique factual matrix and, therefore, the existing law may be applied differently to each case. First, patterns of online harm, as opposed to isolated incidents of online harm, and how this may affect the adjudication thereof is discussed. Second, the importance of context in adjudicating online harms is set out. Next, some of the challenges of addressing online harms due to the nature of the internet and online platforms is discussed. Last, some of the opportunities that technology presents in successfully addressing online harms is explained.

Patterns of harm

It may be that an isolated incident of an online harm does not constitute a justiciable harm. However, the broader pattern of harm against an individual may amount to a justiciable matter. For example, online harassment: when each message is viewed individually, it may be that the conduct in question does not amount to online harassment. However, the pattern of harm may reflect a campaign of online harassment against someone. When viewed holistically, it may be seen that the perpetrator's conduct demonstrates a pattern which, as a whole, amounts to unlawful behaviour. When adjudicating online harms, the judiciary should have a broader look at the harm to identify such patterns.



Additionally, it is important to recognise that these patterns of harm can be perpetrated not just by a single individual but by a group acting like an online gang. This networked behaviour can amplify the impact of harassment and other online harms, making it more insidious and challenging to address. Therefore, when adjudicating online harms, the judiciary should adopt a comprehensive approach, examining the broader context and patterns to identify and address these coordinated campaigns. Including a specific focus on networked behaviour in legal considerations and frameworks can provide a more effective means of tackling such complex and pervasive online harms.

It may also be that a person uses a singular harmful post to incite others to do the same. The original post may not in itself constitute a justiciable harm but, when looking at the ripple effect it has in instigating further harmful conduct from others, it may be that it has created harm beyond that first post. This may then raise the original post to a justiciable online harm.



A case study: online harassment on X

The PHA defines harassment as directly or indirectly engaging in conduct that the respondent knows or ought to know causes harm or inspires the reasonable belief that harm may be caused to the complainant or related person by unreasonably engaging in electronic or any other communication aimed at the complainant or related person.²²¹ This definition does not explicitly provide for a pattern of conduct that may amount to online harassment.

On X, a perpetrator could direct many harmful tweets to a victim or survivor over a period of time. When viewing each individual tweet, it may be that they do not meet the requirements of online harassment in terms of the Act. However, when considered collectively, it may be seen that the perpetrator's conduct shows that they knowingly caused harm and inspired a belief that further harm would be caused. Therefore, their conduct when viewed holistically amounts to online harassment.²²²

Context is key

When adjudicating online harms, the context of the people involved must be considered. The language used, cultural references made, and how these factors relate to the victim and survivor may result in particular actions being more or less harmful to them. Further, this context is necessary to consider in order for the judiciary to be able to holistically determine the intent of the perpetrator, how the act was interpreted by the victim or survivor, and the impact that the online harm had on the victim or survivor.²²³ The online profile of those involved is also relevant for this purpose.

²²¹ See above n215 at section 1.

²²² Power and Associates, 'Securing a protection order against online and offline harms' (2024) (accessible [here](#)).

²²³ P Sekati, 'Context matters: The importance of local nuance in online moderation' (2023) ALT Advisory (accessible [here](#)).



- **Online profile:** The more visible to third parties the online harm in question is, the bigger the impact thereof will be on the victim or survivor, depending on the harm in question. For example, online hate speech will potentially be more harmful if it is seen by a large amount of other people as it could exacerbate feelings of shame etc. Therefore, it is important to examine the reach of the online harm through looking at how many followers or friends the perpetrator and the victim or survivor have, as well as how many other users have interacted with the post containing the online harm. The necessity of such an examination is dependent on the facts – an online harm that has manifested through a direct message may not need to be examined in this way, as only the perpetrator and the victim or survivor would have been able to view it.
- **Language and cultural references:** The language used and cultural references made may impact people differently, depending on their individual context. Further, a person's intention may be established through having regard to their linguistic and cultural context. The degree of harm should also be assessed with this context in mind. The judiciary should therefore examine the post or message in question holistically. Below is a table of some phrases that are illustrative of this point – if not read holistically, the harm that a victim or survivor may experience may be missed, or the intention behind the post may not be established.

Term/phrase	Context	Impact
Necklace/ necklacing	In South Africa, necklacing refers to placing a petrol soaked tyre around someone's neck and setting it on fire. This practice was pronounced during the mid-1980s during the anti-Apartheid resistance. ²²⁴	<ul style="list-style-type: none"> • Harm: For South Africans, the use of the phrase invokes feelings of fear and violence. In contrast, those who do not know the cultural context of the term may take it to refer to jewellery. They would therefore not be impacted by the term in the same way. • Intent: The use of the term in the South African context may clearly show violent intention on the part of the perpetrator.

²²⁴ R Moosage, 'A prose of ambivalence: liberation struggle discourse on necklacing' (2010) Kronos 36 (accessible [here](#)).



<p>Kill the boer, kill the farmer</p>	<p>In 1993, at the memorial of Chris Hani in Cape Town, the president of the ANC Youth League uttered the refrain “kill the boer, kill the farmer’, expressing his anger at Hani’s murder by white conservatives. Post-Apartheid, this phrase has come up many times in various political speeches and protests. The song has a significant relationship to issues of land justice in South Africa.²²⁵</p> <p>In 2022, the Equality Court in <i>Afriforum v Economic Freedom Fighters</i> found that singing the song in the context of the facts at hand did not amount to hate speech as it had to be located within its political context and not taken to be meant literally.²²⁶</p>	<ul style="list-style-type: none"> • Harm: Depending on the people involved, this may be viewed as a threat of violence by the perpetrator. This would possibly invoke feelings of fear to the victim or survivor. Further, this could be viewed by some as a call to violence, which could lead to further harm. • Intent: As set out in the afore mentioned <i>Afriforum</i> case, the context in which this phrase is used is essential to understanding intention. If it is used in a political context, it may be that the perpetrator did not mean it literally. However, if the phrase is directed at an individual not in this context, it may be that the perpetrator’s intention was indeed to incite violence.
---------------------------------------	--	---

- **Emoticons:** The use of emoticons, memes, and other digital expressions can enhance or diminish the impact of online content. Emoticons, for example, can convey emotions and tones that might not be evident in text alone, thereby influencing how a message is perceived. A seemingly innocent message can take on a more sinister tone with the addition of certain emoticons, or vice versa. Similarly, memes, which often rely on cultural references and humour, can either mitigate or aggravate the impact of a message. The judiciary should consider these elements when assessing online harms, as they can provide critical context that influences the severity and intent of the content.

²²⁵ B Brkic, ‘Kill the Boer: a brief history’ (2010) *Daily Maverick* (accessible [here](#)).

²²⁶ *Afriforum v Economic Freedom Fighters and Others* 2022 (6) SA 357 (GJ) at para 99-115 (accessible [here](#)).



Challenges of addressing online harms

Due to the nature of the internet and online social media platforms, there are several challenges in addressing online harms. Some of these challenges – anonymity and pseudonymity and the general difficulty in identifying perpetrators; jurisdictional challenges in the context of cross-border online harms; as well as the challenges associated with gathering evidence from online harms – are discussed here.

- **Anonymity and pseudonymity:** Most online social media platforms allow users to create profiles anonymously or to use a pseudonym. While anonymity and pseudonymity may be seen as a mechanism to allow for freedom of expression and privacy, it may also be misused in the context of online harms. Users are able to hide identifying information on their social media accounts, making it difficult for victims or survivors to identify and hold accountable those who have subjected them to an online harm.²²⁷ While some South African legislation does provide mechanisms for courts to obtain information that would aid in identifying perpetrators,²²⁸ this may lead to protracted and costly court processes for victims and survivors. Further, social media platforms may be reluctant to disclose private information of their users without a court order. Again, this would create difficulty in holding perpetrators accountable.

Identifying the true identity behind an online profile may also be crucial in understanding the context and intent of harmful online behaviour. For example, in the case of *SAHRC v Tracy Zille*, the South African Human Rights Commission received a complaint about a social media account known as “Tracy Zille,” which appeared to be operated by a white woman making disparaging, racist, and harassing comments towards black women.²²⁹ Upon closer inspection and a media expose, it was revealed that the account was actually run by Anthony Mathumba, a local councillor from the Makhado Local Municipality. Mathumba was using the account to generate income through companies he owned and registered, to which social media users were directed. This case highlights the necessity of linking online profiles to phone numbers or URLs to build a clear understanding of who might be behind an anonymous profile. Such linkage helps in accurately identifying the perpetrator and understanding the full context of their actions, ensuring that justice can be appropriately administered.

227 Australia eSafety Commissioner, ‘Anonymity and identity shielding’ (2024)(accessible [here](#)).

228 See, for example, section 4 of the Protection from Harassment Act.

229 South African Human Rights Council ‘Man behind ‘Tracey Zile’ account unmasked, charged with hate speech’ (2020) (accessible [here](#)).





Privacy v accountability: an Australian case study

As mentioned above, anonymity and pseudonymity may at times be an important tool for enabling freedom of expression and privacy on the internet. However, this needs to be balanced when dealing with online harms. In 2020, both Meta and Google reported that they had refused 20% of Australia's law enforcement requests for data held on their platforms.²³⁰ While it is important for platforms to protect personal information, the power that private companies have to disclose, or not, information that could be used to hold perpetrators accountable must be interrogated. Decision making by platforms is often opaque and regulated by their own internal policies, meaning that it is difficult for victims or survivors or even law enforcement to obtain information needed to identify perpetrators of online harms.

- **Jurisdiction:** Due to the nature of social media platforms and the internet more generally, it is possible for a perpetrator to commit an online harm against a victim or survivor in South Africa from outside of the country. In South Africa, a court only has jurisdiction over a matter if the cause of action arose within their jurisdiction, or if the defendant is domiciled or is residence in their jurisdiction. Thus, it may be difficult for our courts to establish jurisdiction over the matter.²³¹ While some countries have attempted to get around this jurisdictional challenge through imposing statutory liability on the platforms themselves, as will be discussed further in the below section, jurisdiction in litigating online harms remains a challenge.
- **Evidence:** On most social media platforms, users are able to delete their posts or messages. Therefore, if a victim or survivor does not screenshot the offending content, it may be difficult to provide evidence to a court and therefore vindicate their rights. While content is difficult to truly and fully remove from the internet once it is posted (due to third-party sharing, caching and archival services)²³² accessing the offending content in question may extremely difficult and ultimately require the cooperation of the social media platform. As has been set out above, this may prove a challenge. However, as will be discussed below, technology make also present opportunities that make accessing evidence easier in some instances.

230 ZD Net, 'Facebook and Google refuse 1 in 5 Australian law enforcement data access requests' (2020) (accessible [here](#)).

231 J Hornle, 'The Jurisdictional Challenge Answered – Enforcement through Gatekeepers on the Internet' *Internet Jurisdiction Law and Practice* (2021) (accessible [here](#)).

232 S Garfinkel, 'How to Delete Regrettable Posts from the Internet' (2012) *MIT Technology Review* (accessible [here](#)).



The opportunities of technology

While technology may pose challenges in dealing with online harms, as discussed above, it also presents certain opportunities that may be leveraged to better adjudicate online harms. This section discusses how technology may be used to access certain types of evidence and the role that experts may play in court proceedings.

Evidence gathering and management

AI and other emergent technology may be used by the judiciary and law enforcement to improve evidence gathering and management. Below is a table of emergent technology and AI tools that could be utilised in this way.

Tool/technology	Application
Automated data collection refers to the use of AI to automatically gather data from multiple sources. ²³³ Automated data collection could be used, for example, to consolidate all of a user's tweets that contain certain phrases.	This tool could be used by prosecution and police investigators to efficiently gather evidence from various online sources, as well as to quickly sort through the data collected. This would make it easier to compile information and evidence relevant to the case at hand. Thus, making investigation and court processes more efficient and thereby improving access to justice.
Natural language processing ("NLP") is a form of AI that uses machine learning to enable computers to understand and interact with human language. ²³⁴	NLP could be used by the judiciary and law enforcement to analyse evidence. For example, it could be used to identify patterns in large amounts of data, as well as collate other meaningful insights such as assessing the true amount of people that may have seen offending conduct.
Blockchain is a decentralised digital database that securely stores records across a network of computers in a manner that is transparent and cannot be tampered with. ²³⁵	Blockchain technology may be used to create secure and immutable records of evidence. Given the vulnerability to manipulation of online evidence, this may become increasingly important to ensure reliability in evidence gathered.

²³³ S Javaid, 'Data Collection Automation: Pros, Cons & 3 Methods' *AIMultiple* (2024)([accessible here](#)).

²³⁴ IBM, 'What is NLP?' (2024)([accessible here](#)).

²³⁵ Investopedia, 'Blockchain Facts: What Is It, How It Works, and How It Can Be Used' (2024)([accessible here](#)).



These technologies can be valuable in ensuring timeous adjudication of online harms through the efficiencies that they can introduce in evidence gathering, analysis and management. It is important to note, however, that technology is not infallible. Especially while these technologies are still developing, it is important to ensure that they are used in a discerning manner, without blind trust. The role of technological experts, discussed below, becomes important in this regard.

Experts

South African the use of experts is not uncommon, with their involvement in matters already regulated by court rules and the law of evidence.²³⁶ As the above discussed technologies, as well as other technology not discussed that may be relevant, are still relatively new, experts in these fields may be useful in court processes and thereby in addressing online harms. Digital forensic experts who have expertise in AI and emergent technology may play an important role in providing clarity and insights in evidence that has been gathered using AI tools. AI and data scientists may aid the judiciary and law enforcement by developing tools to assist in evidence gathering, and by providing analysis into data sets.

Further, experts in online harms and the effects thereof should be utilised in order to provide courts with the proper context of the particular online harm and the effect thereof on the victim or survivor. This will ensure that the online harm and the effects thereof are holistically analysed, ensuring that the nuances thereof are fully ventilated. As this area of law is relatively new, the judiciary would benefit from relying in these experts to create nuanced and up-to-date jurisprudence on contemporary issues.



MMA's expertise in online harms against children

MMA is an example of an expert that could be called upon in matters involving online harms against children. MMA has done a plethora of work around online harms against children, involving advocacy, litigation and skills development. For example, MMA has initiated the Web Rangers, a digital literacy project designed to allow young people to gain critical skills and knowledge around online safety.²³⁷ Due to its work that it has done in this area, MMA could be a valuable expert to courts in providing information about the rights implications that online harms may have on children.

²³⁶ See, for example, rule 38 of the Uniform Rules of Court (accessible [here](#)).

²³⁷ Find out more about the Web Rangers [here](#).



7. Emerging trends

In this section, emerging trends in online harms and their prohibition, both domestically and internationally, will be discussed. While emerging trends and regulations may not be applicable to judicial decision-making yet, this section will provide an overview of things which may soon become relevant. Further, as judges may consider foreign law and must consider international law, this section will provide key developments in foreign and international law. For example, the emerging practice of seeking damages for NCII.

The malicious use of hacking on social media has emerged as a significant trend, largely driven by the rapid growth of digital communication and the increasing sophistication of cybercriminal tactics. With more individuals and businesses utilising social media platforms for communication, marketing, and community engagement, the pool of potential targets for hackers has expanded significantly.²³⁸ This expansion provides cybercriminals with numerous victims to exploit. Furthermore, hackers are employing more advanced techniques, such as phishing scams that leverage social engineering to deceive users. These scams often involve mimicking trusted contacts or reputable organisations, making it easier for hackers to gain unauthorised access to accounts.

The growing use of mobile devices has contributed to this trend. As more users rely on smartphones and tablets, which typically have less robust security than computers, they become attractive targets for hackers seeking access to social media accounts. Additionally, emerging technologies like artificial intelligence and machine learning have enabled cybercriminals to personalise, automate attacks and conduct them at scale. This advancement allows for more targeted and effective hacking campaigns, including automated phishing attempts that can reach a larger audience.

²³⁸ See above n1.



Moreover, social media has become a tool for misinformation, with hackers leveraging these platforms to spread false narratives, manipulate public opinion, and create social discord. These activities can have significant societal impacts, especially during critical events like elections or public health emergencies.²³⁹ The increased anonymity provided by many social media platforms, combined with a lack of stringent regulations regarding user verification and accountability, allows hackers to operate with a degree of impunity. This makes it challenging to trace their activities or hold them accountable.

Financial incentives also drive much of the hacking activity on social media. Whether through identity theft, financial scams, or the sale of stolen data, hackers are motivated by the prospect of profit, leading to an increase in targeted attacks. Given these factors, users, organisations, and policymakers must recognise this emerging trend and take proactive measures to enhance security, educate users about risks, and develop effective legal frameworks to address these challenges. Such awareness can help mitigate the risks associated with hacking and online harms in social media environments.²⁴⁰

Doxing, a term derived from “dropping documents”, refers to the malicious practice of publicly disclosing someone’s private information such as their home address, phone number, or social security number without their consent.²⁴¹ This phenomenon has gained significant traction in recent years, often executed by online mobs aiming to intimidate and instil fear in their targets. The roots of doxing can be traced back to hacker culture in the 1990s, but its implications have evolved significantly, now encompassing issues related to privacy, safety, and ethics in journalism. The misuse of doxing has become alarmingly common, as evidenced by numerous high-profile cases that highlight its damaging impact on individuals’ lives.²⁴² The emotional and psychological toll on victims can be profound, serving as a stark reminder that anyone is susceptible to this form of digital assault.

The Ashley Madison hack in 2015 exemplified the catastrophic consequences of doxing. A group of hackers released sensitive data from the site, which catered to users seeking extramarital affairs.²⁴³ The exposure of personal information led to public humiliation and distress for many users, some of whom faced dire consequences, including suicide. This incident raises serious ethical questions about privacy rights and the responsibilities of platforms to protect user data.

In another notable instance, the doxing of Reddit user Michael Brutsch, known for his controversial posts, led to significant fallout. His identity was revealed, resulting in job loss and public shame. This case illustrates the tension between the desire for online anonymity and the public’s appetite for accountability. While some viewed the exposure as a form of justice, others worried about the chilling effects such actions could have on free speech.²⁴⁴

²³⁹ See above n9.

²⁴⁰ Id.

²⁴¹ S Hinduja, ‘Doxing and Cyberbullying’ *Cyberbullying Research Center* (accessible [here](#)).

²⁴² Id.

²⁴³ A Wallbank, ‘The Ashley Madison data breach: could it happen today?’ (2024) (accessible [here](#)).

²⁴⁴ Id.



Conclusion

This handbook has explored the various types of online harms, particularly the threats posed by online harassment, mis- and disinformation, data protection violations, and the collateral damage caused by content moderation. These harms significantly impede individuals' ability to exercise their constitutional rights, with marginalised groups often bearing the brunt of these negative impacts. As online platforms evolve, new forms of online harassment and violence continue to emerge. Technologies such as AI and deepfakes are enable more sophisticated and anonymous ways to perpetrate harm. Additionally, the increasing use of algorithms to moderate content can lead to over-policing or under-enforcement, particularly affecting vulnerable communities. These trends highlight the ongoing challenges of balancing the protection of digital rights with the mitigation of harm in the virtual space. However, emergent technology also provides various opportunities for combatting online harms.

The role of the judiciary in addressing online harms is critical. Judges must be proactive in understanding and interpreting the complexities of online violence in ways that safeguard human rights while addressing the unique challenges posed by new technologies. Continuous learning and engagement with emerging trends, as well as adopting a human rights-based approach to online harms, is crucial. By fostering transparency, fairness, and accountability, the judiciary can play a key role in ensuring that justice is served in the digital age.



Annexure A: Internet user guide for beginners

Topic	Definition	Tips
Online Harms	Detrimental behaviours and practices occurring in the digital sphere that negatively affect individuals and communities.	Understand the types of online harms and their impact on well-being, privacy, and safety. Report harmful content and seek support when needed.
Online Harassment	Unwanted, aggressive behaviour, such as sending threatening or abusive messages, stalking, or spreading false information.	Block and report harassers. Do not engage with abusive messages. Keep evidence of harassment and seek support from trusted individuals or authorities.
Misinformation and Disinformation	Misinformation is false or misleading information shared without harmful intent; disinformation is deliberately false information to deceive.	Verify information from reliable sources before sharing. Be sceptical of sensational headlines. Use fact-checking websites.
Data Protection and Privacy	Measures to safeguard personal information from unauthorised access and misuse.	Use strong, unique passwords for online accounts. Enable two-factor authentication. Be cautious about sharing personal information online. Review privacy settings regularly.
Content Moderation	The process of monitoring and managing user-generated content to ensure compliance with platform rules and regulations.	Understand the platform's community standards. Use reporting mechanisms to flag inappropriate content. Be aware that content moderation can sometimes result in over-censorship.



Algorithmic Bias and Discrimination	Unfair treatment of individuals or groups by algorithms due to biases in data or design	Be critical of the content suggested by algorithms. Seek diverse perspectives. Advocate for transparency and fairness in algorithm design.
Extremism and Polarisation	The amplification of extreme and divisive content can lead to social division and radicalisation.	Avoid echo chambers by following a variety of sources. Report extremist content. Engage in constructive and respectful discussions.
Cybercrime and Fraud	Illegal activities conducted online, such as hacking, phishing, and financial scams.	Be cautious of suspicious links and emails. Use reputable security software. Monitor your financial accounts regularly for unauthorised transactions.
Psychological and Social Effects	The impact of online harms on mental health and social relationships, such as anxiety, depression, and isolation.	Take breaks from social media. Engage in offline activities. Seek professional help if experiencing severe distress.
WhatsApp	A messaging app used for personal and group communication.	Secure your account with two-step verification. Be cautious of misinformation spread through group chats. Adjust privacy settings to control who can see your profile information.
X (formerly Twitter)	A microblogging platform where users share short messages (tweets).	Use reporting tools to handle abuse. Verify information before sharing. Follow accounts that promote positive and reliable content.



Instagram	A photo and video-sharing platform.	Follow positive accounts. Report harmful content. Be mindful of the potential impact of social media on body image and mental health. Adjust privacy settings to control who can see your posts and stories.
Facebook	A social networking site for sharing updates, photos, and links.	Regularly review and adjust privacy settings. Be cautious about accepting friend requests from strangers. Verify information before sharing.
TikTok	A video-sharing platform known for viral trends.	Use privacy features to control who can see your videos. Be cautious about participating in trends. Report harmful content.
Emoticons/Emojis	Small digital images or icons are used to express emotions, and ideas, or convey messages in text form.	Use emojis thoughtfully; they can help express emotion but may also be misinterpreted. Be mindful of their impact on the tone of your messages. Avoid overuse in formal communication.
Memes	Humorous or cultural images, videos, or text that spread rapidly across the internet, often carrying a particular social or political message.	Memes can entertain, but also be used to reinforce stereotypes or spread misinformation. Be critical of the content shared. Consider whether a meme might be offensive or harmful to others.



Retweeting	Sharing someone else's post or tweet with your own followers on platforms like X (formerly Twitter), amplifies the original message.	Be cautious when retweeting—ensure that the content aligns with your values. Retweeting without understanding the context can unintentionally spread harmful content.
Reports	Reporting is the action of notifying a platform about inappropriate content or behaviour that violates its guidelines or terms of service.	Use the report feature when you encounter harassment, fake news, or harmful content. Be sure to include any relevant information or evidence.
Stories	Short, temporary posts on social media platforms (like Instagram or Facebook) that disappear after 24 hours.	Use stories for personal, timely updates. Be cautious about sharing sensitive information, as stories may be viewed by a broad audience. Review who can view your stories in the privacy settings.
Reels	Short, engaging videos on platforms like Instagram and Facebook, designed to be fun and shareable.	Create and enjoy reels to express creativity. Be mindful of the content you post—ensure it aligns with your values and doesn't harm others. Consider privacy settings when sharing personal content in reels.

