ORIGINAL RESEARCH

Journal of
COMMUNITY SAFETY & WELL-BEING

Digital community management for crime prevention and public safety: Strategies for safer and more inclusive online communities

Sebastian Araghi*, Philip Birch*, Keith Heggart*, John Buchanan*, Hazel Wallace†

ABSTRACT

As social media platforms become central to community communication and engagement, they present both new opportunities and challenges for the prevention, disruption, and reduction of crime through digital public spaces. This article presents findings from a rapid evidence assessment (REA) conducted to inform the Queensland Police Service's Digital Community Safety Champions initiative, focusing on four interrelated areas: de-escalation of online conflicts; dissemination of crime and safety information; best practices for managing crime-focused online communities; and the broader impact of social media on public safety. The REA synthesized evidence from peer-reviewed literature and grey sources published from 2013 up to February 2025, drawing on insights from policing, digital communication, and online community governance. The findings emphasize the importance of context-sensitive moderation strategies grounded in neutrality, timeliness, and discretion. Digital tools that promote deliberative dialogue, such as *TruthMapping*, can support structured engagement and reflection, while post-conflict review strengthens long-term moderation practices. Effective crime communication strategies should combine accuracy, accessibility, visual clarity, and multilingual content to enhance community responsiveness. Best practices for managing online crime communities include establishing clear group norms, safeguarding privacy, building trust through transparency, and avoiding vigilantism through responsible content governance. Finally, while social media offers new avenues for connection and public safety outreach, particularly for vulnerable groups, it also carries risks related to misinformation, radicalization, and surveillance. The article concludes with practical recommendations for moderators, platform designers, and policing stakeholders to help create safer, more ethical, and inclusive digital environments.

Key Words Digital community safety; online conflict de-escalation; community trust; policing; digital ethics; social media.

INTRODUCTION AND BACKGROUND

As digital spaces increasingly serve as forums for community exchange, conflict resolution, and public communication, the role of social media in shaping community safety has become more complex and consequential (Berger & Sklansky, 2023; Forestal, 2022; Tierney, 2013). Platforms such as Facebook, X (formerly Twitter), and community-focused digital forums are now central to how people engage with safety concerns, respond to crime events, and interact with both their neighbours and law enforcement (Hattingh, 2015; Parker & Dodge, 2024). These platforms offer new opportunities for community engagement and information dissemination but also introduce risks, ranging from online harassment and vigilante behaviour to the spread of misinformation, breaches of privacy, and breakdowns in trust (Bikku et al., 2024; Trottier, 2020; Tyagi et al., 2024).

The Queensland Police Service recognized the importance of addressing these challenges proactively through its Digital Intelligence and Community Engagement (DICE) Prevention capability. One of the cornerstone programs of this initiative is the Digital Community Safety Champions project, which seeks to support online community leaders, including moderators and administrators of Facebook neighbourhood and crime groups, in fostering safer and more creditable and positive digital environments. To support this initiative and inform the development of relevant training and policy interventions, the University of Technology Sydney

Correspondence to: Professor Philip Birch, University of Technology Sydney, Faculty of Design & Society, School of International Studies & Education, Building 10, Level 5, 235 Jones Street, Sydney, NSW, 2007, Australia. **E-mail:** Philip.Birch@uts.edu.au

To cite: Araghi, S., Birch, P., Heggart, K., Buchanan, J., & Wallace, H. (2025). Digital community management for crime prevention and public safety: Strategies for safer and more inclusive online communities. *Journal of Community Safety and Well-Being*, 10(3), 138–150. https://doi.org/10.35502/jcswb.478

SGPUBLISHING Published by **SG Publishing Inc.** CSKA Official publication of the **Community Safety Knowledge Alliance**.

conducted a rapid evidence assessment (REA) to support the development of the Digital Community Safety Champions initiative by identifying current evidence and best practices across a range of domains related to digital safety, community leadership, and public engagement in online environments. This article reports on findings from four specific areas of that REA: de-escalation of online conflicts; strategies for communicating crime and safety information in digital spaces; best practices for managing crime-focused online communities; and the broader impact of social media on public safety, especially for marginalized or vulnerable groups. These domains are critically important for ensuring that digital platforms not only remain safe but also actively support inclusive, responsive, and evidence-based community policing strategies. While another article produced through the DICE Prevention initiative focused on digital literacy and moderation practices (see Araghi et al., under review), emerging challenges related to online conflict escalation, crime and safety information dissemination, community management, and the wider societal impacts of social media on public safety were also revealed.

METHODOLOGY

The study presented in this article draws on a REA designed to identify and synthesize best practices relevant to digital community safety, communication, and engagement. The REA method was selected for its ability to provide a structured, transparent, and timely review of available literature to inform practical interventions in fast-evolving digital environments. While less exhaustive than a full systematic review, the REA offers more methodological rigour than a traditional narrative literature review and is well suited to informing time-sensitive, policy-relevant decisions.

The REA was guided by the following research questions:

1. What are the best practices for digital community safety?
2. How can misinformation and online scams be effectively identified and combated?
3. What are the effective strategies for online community moderation?
4. How can conflicts in online communities be de-escalated?
5. What are the strategies for disseminating crime and safety information online?
6. What are the best practices for managing online communities, especially those focused on neighbourhood and crime?
7. How can digital literacy and education improve community safety?
8. What is the impact of social media on public safety?

The results reported in this article draw on four of those eight questions guiding the wider REA,[1] including the following:

1. How can conflicts in online communities be de-escalated?
2. What are the strategies for disseminating crime and safety information online?
3. What are the best practices for managing online communities, especially those focused on neighbourhood and crime?
4. What is the impact of social media on public safety?

These questions were selected based on their relevance to the DICE Prevention project's expanded scope and their potential to inform proactive, community-centred approaches to digital conflict management, communication, and governance. In essence, this paper centres on exploring *why* trust, de-escalation, communication clarity, and community values matter.

SEARCH STRATEGY

The search terms governing the REA were "digital community safety initiatives"; "online misinformation detection"; "community moderation strategies"; "conflict de-escalation online"; "crime prevention information online"; "online community management best practices"; "digital literacy for community safety"; and "social media public safety impact."

The specific keyword-based search strategy used for each of the four questions presented in this article included phrases such as "online conflict de-escalation," "digital community engagement," "crime information dissemination," "neighbourhood digital communities," "social media public safety," and "digital trust and communication". Where search results yielded large volumes of material (e.g., for "social media and public safety"), filters such as "Facebook" or "online communities" were added to focus results.

The inclusion criteria were as follows:

- Studies and reports published in the last 10 years (2013 up to February 2025).
- Materials that address one or more of the defined research questions.
- Peer-reviewed articles, governmental reports, and reputable organizational publications.

Exclusion criteria comprised the following:

- Irrelevant studies that do not address the specified topics.
- Publications not available in full text.
- Non-English-language materials (unless translations are available).

While the REA was designed to be as inclusive as possible within time and resource constraints, the exclusion of non-English sources (unless translated) was necessary. Similarly, although grey literature was included when published by reputable organizations (e.g., government bodies, international NGOs), lower-tier or unverified grey sources (such as unmoderated forums or unarchived blog posts) were excluded to ensure the credibility of findings. Furthermore, while platforms such as Facebook and X (formerly Twitter) feature prominently in the literature, we acknowledge that this reflects current publication trends rather than an exhaustive survey of all social media ecosys-

[1]Questions 1, 2, 3, and 7 of the REA are presented in Araghi, S., Birch, P., Heggart, K., Buchanan, J., & Wallace, H. (manuscript submitted for publication). *Digital community safety and crime prevention: Insights from a rapid evidence assessment of online moderation practices.*

tems. Future research could usefully extend this work by examining decentralized, encrypted, or platform-agnostic networks (e.g., Mastodon, Discord, Reddit, or peer-to-peer forums), which may provide unique insights into digital conflict and safety practices.

The search was conducted across multiple databases including Scopus, PubMed, and Web of Science, as well as key organizational repositories such as those maintained by the Australian Institute of Criminology and the Office of the eSafety Commissioner. Each included source was recorded in a review database and coded for citation details, methodological approach, key findings, and relevance to the guiding questions. Articles were excluded at the abstract level if they focused on vertical digital communication (e.g., individual consultations with professionals) rather than horizontal community-based interactions. Full-text exclusions were based on relevance, accessibility, or redundancy.

The number of articles used for the REA is set out in Table I, including the number of articles used to inform the analysis presented in this article.

This article presents a thematic synthesis of the retained literature, organized around the four guiding questions. The findings inform practical guidance for police, moderators, and digital community leaders seeking to create safer, more ethical, and more inclusive digital spaces.

FINDINGS

The following findings presented below provide a synthesis which focuses on questions 4, 5, 6, and 8 that centred on (1) de-escalating conflicts in online communities, (2) disseminating crime and safety information, (3) managing online communities (particularly those focused on neighbourhood and crime), and (4) understanding the broader impact of social media on public safety. Each theme is explored with attention to evidence-based strategies, emerging challenges, and practical insights that can inform the design of safer, more inclusive digital communities. The discussion draws on diverse disciplinary sources, including digital communication, policing, online moderation, and public safety, with a view to providing actionable guidance for moderators, law enforcement, and community leaders alike.

De-escalating Conflicts in Online Communities

Conflict within digital communities is different to normal conflict as it is often witnessed by a large audience and can be continued at any time (Marwick & Boyd, 2014; Peter & Valkenburg, 2013). Moreover, the absence of any physical cues can make it more difficult to effectively resolve conflict with another person (Berger, 2013). As such, it is important to have a procedure to de-escalate conflicts successfully as a moderator, including mediating and facilitating resolution in a timely manner; encouraging constructive dialogue while remaining neutral and calm; enforcing community guidelines if necessary; utilizing automated tools if they are beneficial; and reviewing and reflecting your decisions at a later point to determine if any improvements or changes could have been/need to be made.

TABLE I REA article screening process

Criterion	Total	Remaining Total
Selected for initial screening	532	Not applicable
Rejected at abstract level/repeats	242	290
Rejected at full-article level	119	171
Article reports on questions 4, 5, 6, and 8	171	84

REA = rapid evidence assessment.

De-escalating online conflict and facilitating resolution can be achieved through means such as negotiation, mediation, engagement, clarification of an organization's *raison d'être*, problem-solving orientation, and consistency in approach (Milofsky et al., 2017). Milofsky et al. (2017) add that this requires training (which emphasizes the importance of building a strong moderation team). Moreover, consistency of approach may be problematic, given the dynamic, sometimes volatile, nature of online communication (Singh, 2013). De-escalation also requires open communication, to promote inclusion, which may serve to counter hostilities borne of feelings of isolation, real or imagined (Hirblinger, 2020). While Hirblinger's context is international relations, it surely applies equally at smaller scales. As the International Crisis Group (2020) points out (in the context of Myanmar), the party that is in, or sees itself as, the ascendancy, may be less willing to de-escalate conflict, frustrating resolution through negotiation.

Moderators play a useful role in conflict de-escalation, and if they are available at all hours of the day and trained well, they should be able to stop the conflict almost as immediately as it is being started, acknowledging the issue in a timely fashion and starting steps to resolve it (Dineva & Daunt, 2023). Lane and Stuart (2022) challenge the notion that social media conflagrate violence, arguing that, through observation and mediation, users can equally intervene and de-escalate tensions. This can occur through intervention of a third person (a moderator), or by self-regulation, to avoid conflict (see also Kuan-Ming (2024), above). If moderators act as a neutral third person who is impartial toward the parties, issues, and interests in question, it should help them to facilitate the parties resolving the dispute in a reasonable amount of time (Moore, 2014). For example, moderators can take different approaches to de-escalate conflict. They may choose to pause content by temporarily disabling comments or posts if a discussion is spiralling out of control, allowing time to address the issue privately with those involved. Alternatively, they can offer constructive solutions, such as encouraging participants to use tools like "ConsiderIt" or "TruthMapping17," which help users calm down, reflect, and better understand different perspectives. These tools will be explored in more detail later.

However, other literature does highlight the important role that the media has in this regard. For instance, Morah and Oladokun (2023) examined media reporting of terrorism, wherein the stakes are high, in Nigeria. They contend that "the social responsibility performance credential of the reporter can help build a nation or be a factor of the collapse of social control" (p. 346). In essence, they believe that if reporters act responsibly, their work could also contribute to informed discourse and unity, preventing potential distrust and panic from occurring at all.

Next, after the initial stage of addressing the conflict and starting the mediation process has occurred, moderators should encourage constructive dialogue while remaining neutral and calm themselves. As mentioned above in Moore (2014), remaining neutral and calm is a key component to successfully moderating conflict as it lessens the likelihood of you furthering the conflict through your involvement, and instead gives you a better chance of facilitating resolution. Moreover, it is important that once you begin to communicate with those involved in the conflict that you move the discussion to private messages as it has been shown that intervening on social media while there is an audience during the conflict only amplifies the problem as it is hyper-visible (Elsaesser et al., 2019). While Elsaesser et al. (2019) found that the presence of an audience serves as a protective factor against online conflicts escalating into real-world violence, they also highlighted its negative impact in online spaces, where the audience is more likely to exacerbate the conflict.

Once a calm and neutral tone has been established and the discussion has been moved to private messaging, the moderator should aim to reframe the conversation, guiding individuals toward problem-solving rather than personal attacks. Online platforms have proven highly effective in mobilizing dissent and empowering marginalized voices (Bharati et al., 2022; Leman-Langlois et al., 2024; Türker & Gök, 2024). Conflict within online communities can sometimes be disruptive, but it can also be productive when it challenges power structures, promotes democracy, and fosters innovation. However, if left unchecked, it can lead to misinformation and manipulation, where a new dominant group replaces an old one without necessarily improving fairness or inclusivity. While mobilizing allies for a worthwhile cause can be advantageous, excessive groupthink may stifle innovation rather than encourage it. This highlights the dual nature of online conflict – both as a force for positive change and a potential source of harm. As part of addressing this balance, moderators should consider the dynamics of group identity, leadership structures, and individual roles within the community (Milofsky et al., 2017) to ensure that discussions remain constructive rather than destructive.

Thus, to ensure that conflict remains constructive rather than disruptive, moderators can introduce tools that help users engage with differing perspectives in a structured and reflective manner. Platforms like "ConsiderIt" or "TruthMapping17" provide frameworks that encourage thoughtful deliberation, allowing individuals to process disagreements without resorting to hostility or escalation. For "ConsiderIt," the platform invites users to think about the trade-offs of a proposed action by creating a pro/con list. This list creation is augmented by allowing users to input the points that have also been contributed by others. By doing so, this can allow users to gather insights into the considerations of the people on the other side of the conflict, potentially identifying unexpected common ground (Jhaver et al., 2018; Kriplean et al., 2013). Additionally, Jhaver et al. (2018) highlights how the platform's focus on personal deliberation, as opposed to direct discussion with others, reduces the opportunities for conflicts.

Similarly, "TruthMapping17" allows users to collect and organize ideas, constructively test those ideas, and promote reasoning-based discourse. This tool structures conversations using argument maps, critiques, and rebuttals. It invites users to break down a topic into its component parts – assumptions and conclusions – and create a node for each part, so that the hidden assumptions are made explicit. All critiques are directed against specific nodes so that any attempts at digression are apparent. Only the original arguer can modify the map, but any user can provide their feedback by adding a critique to any assumption or conclusion or by responding to a previously posted critique with a rebuttal. By doing so, it allows the user to gather a multitude of perspectives on the topic, which may align or clash with their own personal views, encouraging them to reconsider their perspective rather than attacking another for having a different one (Jhaver et al., 2018).

Thus, these are two solutions that moderators can propose to facilitate resolution. However, their effectiveness depends on the user's willingness to engage honestly with the material rather than manipulate or misrepresent the information, meaning they may not always succeed. In cases where this approach fails to reframe the conversation, moderators can take a firmer stance by reinforcing community guidelines on respectful behaviour (as discussed in question 3). They can remind participants of the expectations for constructive discussion and the potential consequences of violating these rules where necessary.

If the individual continues to escalate the situation or act disrespectfully, it is crucial to enforce community guidelines. Moderators have the option to issue warnings, mute users, or, in more severe cases, impose bans. The severity of the misconduct should guide the moderator's response, with muting often serving as a fair initial step to avoid disproportionate punishment (Jhaver et al., 2018). As Jhaver et al. (2018) explains, doing so would create a hybrid moderation mechanism whereby it ameliorates the risk of individuals being permanently blocked for relatively minor infractions. Instead, "blocklists" can be reserved for more serious violations, such as harassment. Consequently, this also shows the importance of human moderation as opposed to just using artificial intelligence (AI)-driven tools to moderate, as it allows for a distinction when the situation is nuanced. Furthermore, having human moderators allows for a clearer communication of why moderation actions were taken, which Jhaver et al. (2018) suggests can help de-escalate rancour from the participants of online communities, particularly those who are blocked.

While not as commonly used, automated tools such as big data and its predictive capacity can be used to pre-empt conflict through means including early warning and rapid feedback (Letouzé et al., 2013). Such data might also serve to devise appropriate interventions for likely offenders, and to identify possible suspects following a crime (Perry et al., 2013). However, Lee et al. (2021) point out the tension between big data collection and liberty. Lee et al. (2021) also point out that big data analytics are not free from subjectivity. As De Bruyn (2021) notes, online communication alone cannot provide a full picture of users. As such, while these tools can be valuable, it is essential to consider the ethical concerns raised above, particularly regarding the risks of predicting crime – namely, issues of accuracy, potential bias, and the trade-off between security and individual privacy (MacCarthy, 2023; Pauwels, 2020; Perry et al., 2013).

Finally, it is important that, months or years in the aftermath of the conflict, moderators review and reflect on the outcome, determining whether different actions should have been taken. By analyzing what triggered the conflict and how it was handled, they may be able to determine an alternate course of resolution that could have been taken, or whether the community guidelines need to be revised. In doing so, this more effectively prepares them to handle similar conflicts in the future, allowing for more effective conflict de-escalation and a stronger digital community.

Strategies for Disseminating Crime and Safety Information Online

While specific data on the best strategies for disseminating crime-related information is limited, we can extrapolate insights from fields like public health, disaster response, and general safety to build an effective framework for online crime and safety communication. The first and most crucial step is using trusted, widely accessible platforms, as the primary challenge lies in ensuring knowledge of current best practices reaches those who need it most (Chan et al., 2020). These platforms should also facilitate meaningful community engagement, encouraging dialogue and participation. Additionally, clear and concise messaging is essential to ensure accessibility, while technology should be leveraged to reach key overrepresented demographics and enhance accessibility through tools like mobile apps. Finally, maintaining accuracy and credibility throughout is vital to prevent misinformation and foster trust in the information being shared.

As we progress into an increasingly digital world, utilizing social media to disseminate information has been shown to be a speedier alternative (Ng et al., 2020), particularly with surveys showing that at least 74% of individuals are using social media platforms such as Twitter to generate their daily news every day (Rosenstiel et al., 2015). This is because the paths for, and rate of dissemination of traditional scholarly publications, static websites, and even emails, have been proven to be slow (Brownson et al., 2018), whereas studies have proven social media to be a significantly stronger method of dissemination. For instance, Chan et al. (2020) conducted a study whereby they disseminated an infographic via Twitter (now called X) and WeChat (which is a Chinese social media app), while also sharing the same information on a government website and more traditional methods of dissemination. In this study, not only did they find that the social media had received approximately 760% more page views (63,440 compared to 8,614) than traditional methods, but they also found that sharing it through social media allowed it to be redistributed in other areas, and languages, where necessary, with locally facilitated translations occurring in Italian, Portuguese, French, German, and countless more.

However, as outlined by Rodilosso (2024), there are risks to social media dissemination, for instance, the algorithm-driven filter bubbles that selectively display information based on user preferences. Additionally, social media can foster engagement with biased knowledge within echo chambers, spaces where like-minded individuals interact, resulting in negative consequences (Conway et al., 2019), with Chan et al. (2020) highlighting the importance of free open-access educational material that distills key information in a clear format being used in conjunction with social media, as well as traditional methods. Moreover, utilizing social media allows for better community engagement to be fostered, enhancing dissemination efforts. This can include interactive Q&A sessions where community members ask questions about crime and safety, promoting tools for reporting suspicious activity such as tip hotlines or web forms, and conducting surveys to collect community input on safety concerns and preferred communication methods. These approaches encourage active participation and help ensure that safety messaging resonates with the audience.

Additionally, even if just social media is used, Zhu et al. (2018) proposed a model for information with constant update which would mitigate some of the risks of social media dissemination. DMCU, as they coined it, determined the priority of related information by modelling the general behaviours of the user. Moreover, it ensures that updated information always has higher priority than the original, because it generally contains more valuable content (and can supersede the original), which is important as authors and disseminators on social media tend to modify information if they learn it is false after they have already shared it (Simon et al., 2015). Thus, the information update process within DMCU is continuous and successive.

Even more, the DMCU model also considers that environmental factors such as the source information node, dissemination probability of information, the information update time delay, and negative feedback on information are all factors that affect the process of information dissemination (Zhu et al., 2018). These environmental factors highlight the complexities of online information dissemination, demonstrating that simply posting safety updates is not enough. Ensuring that information comes from credible sources, reaches the intended audience efficiently, and is regularly updated without excessive delay can help mitigate misinformation. Moreover, by accounting for how users interact with content – whether by amplifying, questioning, or rejecting it – authorities can refine their communication strategies to maximize engagement and trust. By integrating these considerations into crime and safety messaging, online platforms can improve public awareness and response effectiveness.

Once the platform is chosen, it is important that all information disseminated has a clear and concise meaning, avoiding jargon and ensuring that it is understandable to the average person. As outlined by Freiden (2013), communication is an essential component of dissemination because it supports behaviour change, political commitment, and social norms that shape the context for public discourse. Since the language within dissemination is important, guides for how to structure dissemination can and should be utilized as they acknowledge the importance of not just identifying an audience but also delivering clear and actionable information to them in language they understand (Baur & Prue, 2014). For instance, the Centers for Disease Control and Prevention (CDC) developed an index in the United States that aims to make materials more understandable, with respondents able to more clearly identify the intended main message and understand the words and numbers in the materials better for 9/10 questions (Baur & Prue, 2014). Thus, while the CDC's index was not originally designed for crime-related information, applying its principles to crime and safety

messaging could enhance comprehension, making it more likely that readers will clearly understand and act upon the information provided.

As well as this, another aspect that can help the user to understand information is visualization, such as infographics, maps, or videos. For example, well-designed infographics have the potential to provide concise and practical information while requiring a lower cognitive load and being preferred by the reader (Thoma et al., 2018). Moreover, they aid knowledge translation by increasing information retention according to the cognitive load theory and dual coding theory (Martin et al., 2019). Thus, making infographics easily accessible, engaging, reusable, and modifiable to fit local needs and user requirements makes the information more likely to be successfully disseminated, allowing for crime and safety information to be more widely received and retained (Chan et al., 2020). Moreover, as Chan et al. (2020) showed, completing translations of the information within a 10-day period allowed for diverse audiences from all backgrounds to have access, while also allowing for context-specific modifications (e.g., Italy requesting that a section of an infographic talk about "double gloving" during COVID-19). Thus, by combining infographics and multilingual content, Chan et al. (2020) was able to disseminate their information to 63,000 people (as referenced above).

Leveraging technology is another effective strategy that can assist in disseminating crime and safety information due to the multitude of options it provides. For example, geotargeting is crucial as it can use alerts to reach individuals in affected areas promptly. By using location-based notifications, authorities can provide real-time updates (using mobile apps and even push notifications) on crimes or any hazards, improving public awareness and response times. For example, the Philippines' Project NOAH was developed in 2012 to mitigate the impact of natural disasters by using a Web-GIS tool that provided early warnings about impending floods, giving at-risk communities up to 6 hours of advance notice (Lagmay et al., 2017). A similar model could be applied to crime prevention, where geotargeted alerts inform residents about nearby incidents, such as active threats or missing persons. However, as Roystonn et al. (2023) highlight, digital platforms must be culturally responsive and accessible to avoid disempowering communities or excluding those without digital access. By integrating geotargeting with inclusive communication strategies, authorities can enhance public safety while ensuring equitable access to critical information.

Finally, throughout the process of disseminating crime and safety information, it is crucial to ensure that the content is both accurate and credible. This can be achieved by fact-checking all information before dissemination to prevent the spread of misinformation, which can undermine public trust. Additionally, referencing official sources such as law enforcement or government agencies is essential for building trust in the information shared. Equally important is providing clear attribution to ensure transparency, as clearly stating who is providing the information – such as the police department, the Australian Institute of Criminology, or the Australian Bureau of Statistics – will reinforce the credibility of the message. Importantly, this step must be emphasized throughout the entirety of the dissemination process to ensure that the information remains reliable, trustworthy, and effective in fostering safety.

Best Practices for Managing Online Communities, Especially Those Focused on Neighbourhood and Crime

Effectively managing online communities, particularly those centred on neighbourhood safety and crime, requires a strategic approach that balances security, engagement, and trust. Best practices include establishing clear community guidelines to set expectations, fostering transparency to build trust, and implementing strong moderation strategies to maintain respectful discussions. Additionally, safeguarding the privacy of all individuals – including members, victims, witnesses, and even suspected perpetrators – is crucial. Moreover, a sense of community should be fostered through encouragement and recognizing contributions, making it easier to address conflicts and misinformation swiftly to ensure a safe and reliable space for all participants.

First and foremost, establishing clear community guidelines by defining the purpose of the digital community is crucial as users will not join a community if they do not understand and relate with its mission. As outlined by Sun et al. (2014), there are four main factors that influence people to join digital communities: environmental factors that are not related to the user; individual factors referring to the personal characteristics of the users; commitment factors; and quality requirement factors. Thus, the ability of the user to see accurate, well-moderated and reliable information is especially important for enticing users to join and participate. In a crime prevention-focused group where potential misinformation can spread easily (as outlined in question 2), if the community is filled with speculation, false crime reports, or fearmongering, some users may leave due to a lack of trust in the information provided. However, if the community has a clearly outlined well-defined purpose such as sharing verified crime reports and promoting neighbourhood safety initiatives, it aligns with users' quality expectations and makes them more likely to join (Sun et al., 2014).

Moreover, another component of establishing clear community guidelines is setting rules for engagement, including respectfully communicating, which, as referenced above, is crucial to prevent the exclusion and ostracism which can occur from online community engagement (Hembroff et al., 2020). Also, it should be made clear what content is acceptable in an attempt for transparency, so that user frustration will be reduced, and instances of uncivil behaviour will be lowered in the event that a moderator needs to mute or ban a user for posting unacceptable content (Hernandez-Bocanegra & Ziegler, 2021).

Finally, it should also be specific that there is zero tolerance for abuse. As discussed in other sections, networks rely heavily on interpersonal connections and a shared purpose, often referred to as" consortium-building" (Qadir et al., 2024, p. 1). Ensuring that abusive behaviour is swiftly addressed is essential to preserving this sense of community. For smaller groups, issues like trolling can often be managed directly by moderators (Dineva & Breitsohl, 2022). However, on larger platforms, moderation must extend beyond individual cases to address systemic factors that contribute to "deviant online behaviour" (p. 300). This requires a multi-tiered approach,

including clear policies at the community level, proactive moderation strategies, and, as Dineva and Breitsohl (2022) suggest, interventions at the organizational level to reinforce a culture of accountability and safety.

While clear community guidelines provide a foundational framework, transparency and trust are equally essential in fostering a safe and credible online space. Users need to feel confident that the information they receive is reliable and that it is safe to speak, post, or comment. By clearly identifying moderators and administrators, accountability can be established, allowing members to see who is responsible for decision-making within the community. Furthermore, as previously discussed, openly communicating moderation actions builds trust, enhances perceptions of fairness, and reassures participants that shared information is accurate (Grimmelikhuijsen & Meijer, 2015; Hernandez-Bocanegra & Ziegler, 2021).

Importantly, trust should also be built by ensuring that all information sourced and provided is credible, sharing accurate information such as verified crime alerts and resources from trusted sources. This is especially important for digital communities as their reputation is dependent on them being transparent and trustworthy, and reputation is considered one of the five motivational factors convincing users to join online communities (Sun et al., 2014). Consequently, bilateral communication between citizens and law enforcement plays a crucial role in ensuring that crime-related information shared within online communities is accurate and credible. Sachdeva and Kumaraguru (2015) examined an online crime prevention network in Bangalore, India, which not only provided a platform for reporting crime but also allowed users to seek guidance on appropriate services to contact in various situations. The network facilitated direct engagement with police, who responded to queries and feedback on their performance. Notably, most users reported satisfaction with the responses they received, reinforcing trust in the network's reliability. This responsiveness enhances the credibility of the information shared and fosters confidence in the platform as a legitimate resource for crime-related concerns. However, in contexts where trust in law enforcement is low (Ho & Cho, 2017; Sigsworth, 2019), ensuring transparency in communication remains particularly important for maintaining community engagement and trust. Therefore, this provides evidence that cooperation with law enforcement is one of the best practices for providing credible information, and therefore, ensuring transparency and trust within a crime-related online community.

While moderation practices have been extensively explored already, it is particularly useful to apply these practices to the specific lens of a crime and neighbourhood-related digital community, where moderation can often be essential to maintain civility. Online interactions are often more prone to incivility than face-to-face conversations due to the disinhibiting effects of anonymity and deindividuation (Lowry et al., 2016). Without proper monitoring, this can escalate into cyberbullying or harassment, as perpetrators may not fully grasp the impact of their actions. Lowry et al. (2017) highlight the control imbalance between victims and aggressors in online spaces, reinforcing the need for proactive moderation. However, with trained moderators who can swiftly address inappropriate content and enforce community rules, online platforms can foster positive engagement. Additionally, research by Ghouri et al. (2020) demonstrates that digital spaces, when moderated effectively, can also be leveraged to promote peace and constructive dialogue, demonstrating the potential of well-moderated communities to encourage respectful and productive discussions.

While most of these other strategies have been more generalizable to all digital communities, protecting privacy is something that is particularly relevant for crime-related communities due to the sensitive nature of the content that is often being shared and explored. For instance, it is important to secure the platform so that unauthorized access or breaches about potentially sensitive information can be identified and educate members on how to protect personal information when reporting crimes or discussing sensitive issues. Moreover, it is important to anonymize sensitive reports to ensure that identifiable details of victims, witnesses, and even suspected perpetrators are not disclosed.

While it is widely accepted to protect the identities of victims and witnesses to respect their privacy and rights, this same level of protection is often lacking for suspected perpetrators. For example, since 2013, there has been a shift within the Victorian police to identify suspects not just after they are convicted, but during their criminal investigations (McMahon, n.d.). As McMahon highlights, despite Australia having a theoretical "presumption of innocence," this presumption is fragile, as many people correlate being charged or prosecuted with also being guilty. However, insufficient consideration is given to the potential damage to the reputation of individuals who were wrongly suspected of committing a crime but still faced significant abuse. For example, in the case of Richard Jewell, an American security guard, he lost his job and was subjected to harassment, public scorn, and social ostracism, despite later being revealed to have helped people to safety rather than being the terrorist he was initially accused of being (McMahon, n.d.).

Furthermore, this practice of non-identification is extremely important to follow for crime-related digital communities, as there have been many proven cases of vigilantism in response to the general community having access to victims' names. For example, in the United States alone, 279 incidents of vigilantism (featuring 427 separate vigilantes) took place between 1983 and 2015 whereby a vigilante (defined as a private citizen acting without legal authority) targeted, attacked, and sometimes even murdered an individual who had been previously accused (real or false), arrested, convicted, or registered for a sex offence (Cubellis et al., 2019). While Cubellis et al. (2019) highlights that sex offences tend to provoke strong emotional responses, sometimes leading third parties to take violent action, this trend underscores the broader risk of naming suspected perpetrators (or potentially even perpetrators) in online communities. Sharing these names can fuel unlawful retaliation, undermining the legal process and community safety, while also potentially punishing the wrong person in the cases where the offender is incorrectly identified.

Regarding fostering a sense of community, a strong sense of mission (which is established above) and purpose is essential for fostering positive interactions within an online community. Tim et al. (2014) highlight the role of advocacy in "catalyzing civic engagement," demonstrating how shared

goals can bring members together. They explore the concept of "boundary objects" – tools or ideas that are flexible enough to meet different users' needs while remaining stable enough to create a common understanding over time. This shared meaning emerges through collaboration between network organizers and users, reinforcing a sense of belonging. To foster engagement, network organizers must also communicate in ways that resonate with both current and prospective members (Cheong et al., 2023; Fourie et al., 2022). Similarly, external frameworks, such as the Queensland Government's online community charter, can provide clear expectations that guide constructive participation and reinforce community cohesion (Sullivan et al., 2019).

Moreover, it should be common practice to recognize contributions by thanking members who actively participate, with Van Mierlo (2014) underscoring how the top 1% of users tend to account for 73.6% of post creations and the next 9% account for 24.7%. As such, thanking these members is likely to foster engagement by serving as a reward for a job well done, encouraging them to post and report more in the future. However, Edelmann (2013) also highlights how lurking is also a normal, active, and participative and valuable form of online behaviour, showing that it may be beneficial to thank the seemingly inactive members as well, as they could be utilizing the crime-related information for public good. Finally, a sense of community could also be created by organizing offline events to strengthen offline connections.

The Impact of Social Media on Public Safety

Digital technologies have been deployed to enhance security. But this raises questions such as security for whom, against whom, for what purposes, and in whose interests (Kleinig et al., 2011)[2]. To play with Kleinig et al.'s chapter title, what motives might be underlying deployment of digital technologies by those in a position to do so? The same questions apply to social media – who benefits (if anyone), who is targeted, and whose interests are being served?

One group that clearly benefits from social media in a public safety context is vulnerable individuals. For many, digital platforms provide essential support systems, avenues for advocacy, and tools for crime prevention. While risks such as harassment and exploitation exist, social media also enables marginalized groups to access vital resources, report crimes, and connect with communities that offer protection and empowerment. Moreover, social media's capacity for connecting people across distance is a vital tool for creating supportive communities whereby individuals who face harassment or bullying can feel heard, supported, and empowered to act.

However, for these vulnerable groups such as women, people of colour, minors, people within the LGBTQIA+ community, those experiencing domestic and/or sexual violence online harassment can often amplify their marginalization, making it essential to raise awareness about the dangers they encounter. For example, research by Miller and Lewis (2023) highlights the disproportionate levels of harassment directed at women and transgender individuals, as well as people of colour and those who publicly display their faith through clothing or symbols. Such harassment can create an unsafe environment, particularly when vulnerable individuals are subjected to targeted abuse.

Minors are especially at risk, as they may face bullying, predation, and exposure to harmful content. Schwartz et al. (2016) explored the safety issues in schools, emphasizing that girls are disproportionately affected by online harassment. Furthermore, the elderly are also more susceptible to online exploitation, including scams and abuse. Recognizing these vulnerabilities, the Australian Government has recently passed legislation to exclude minors under the age of 16 years from some social media platforms to protect young people from the dangers posed by online spaces.

Importantly, social media platforms can provide opportunities in response to these risks, such as peer support and advocacy. For instance, research conducted by Do et al. (2023) on Korean adolescents with attention-deficit hyperactivity disorder found that online spaces could offer a "safe space" for sharing sensitive issues that young people might be reluctant to discuss with their parents, such as self-harm and suicidal thoughts. Thus, this is an example of social media providing marginalized and vulnerable individuals with the opportunity to share experiences, find solidarity, and access support networks, which is crucial for their mental and emotional well-being.

Another positive impact that social media has on public safety is through community engagement and awareness, as law enforcement is increasingly using social media platforms to build trust with communities, promote public safety, and even disseminate critical information. However, as established prior, clear communication is especially vital when engaging with all communities, but particularly diverse communities, as cultural differences can impact how messages are received and understood (Bailey, 2015; James & Wilfred, 2024). Moreover, as covered above, efforts to foster positive relationships between police and the community are crucial for enhancing public safety, and social media can serve as a bridge for these exchanges. Transparency in communication, such as sharing updates about ongoing investigations or safety tips, helps build stronger relationships and promotes collaboration, which is essential for creating safer neighbourhoods.

Furthermore, the risk of digital harassment, particularly for marginalized groups, also points to the need for law enforcement and community organizations to use social media as a tool for awareness campaigns. These campaigns can raise awareness about specific dangers faced by vulnerable groups, such as online harassment, cyberbullying, or scams. Consequently, public campaigns on how to identify and report such issues also empower the community to act proactively, ensuring a safer online environment for everyone.

However, while there are many positive benefits to social media, there are also significant negative impacts that social media has on public safety, including spreading misinformation and panic (including exploitation), and an increased risk of harm, both digital and physical. In the case of spreading misinformation, since Facebook and Instagram have stopped taking accountability for misinformation and spam content (McMahon et al., 2025)[3], and the Australian Government

[2]Source not included in the REA but added into the narrative for context.

[3]Source not included in the REA but added into the narrative for context.

has abandoned its attempts to pass a law requiring online platforms to monitor misinformation (Evans, 2024)[4], there is a significantly higher likelihood of coming across fake reports or false emergency alerts that can cause unsafe decisions or overreactions. Moreover, Nabiullina (2021) found that even after teaching a class about the risks of disseminating disinformation, 14.3% of students answered "sometimes" and 3.9% of students answered "yes" in a survey where they were asked if they would still disseminate any information without verifying its reliability. This suggests that some individuals on social media remain indifferent to the accuracy of the content they distribute. Given this challenge, Nabiullina (2021) proposes that an updated deep neural network would be needed to enhance the detection of false information beyond current capabilities.

Additionally, another instance where social media can spread panic is through manipulation and exploitation, whereby extremist groups can exploit social media for propaganda purposes, recruiting individuals, or planning illegal activities. For instance, lone wolf terrorists often undergo a process of self-radicalization, where media influence drives them to act in alignment with terrorist organizations despite having no direct ties to these groups (Cohen, 2013). Moreover, this problem is becoming significant enough that Byman (2017) argues that lone wolves should be completely isolated from being able to talk to like-minded people, ensuring social media companies like Facebook and X tighten restrictions so that groups such as ISIS or other extremist communities cannot have a platform. However, at this date there are no restrictions preventing this, allowing extremist groups to be able to radicalize users without even having direct contact and therefore serving as an extremely significant risk to public safety.

Another negative impact of social media on public safety is that it increases the risk of harm, through hate speech and cyberbullying online, and physical harm, through the tracking of locations and targeted attacks. As referenced in question 1 through Saleem et al. (2017), hate speech currently must be removed by moderators on platforms such as Facebook, making it extremely harder to identify and remove. As such, this is why instances of mass cyberbullying and hate speech are occurring, such as the 1,634 Malaysians who reported instances of online victimization and hate speech (Marret & Choo, 2017).

Finally, the ability to track locations and target attacks through social media is a significant concern, with 18% of 4,800 respondents in the Australian eSafety Commissioner's 2022 survey reporting being tracked electronically without consent, and 16% experiencing online threats of in-person harm or abuse (AIHW, 2024). Moreover, if users access a GeoSN application through their mobile device – such as checking in on Facebook – their personal information may be used for undesirable purposes, compromising their privacy (Alrayes & Alia, 2014). This breach can involve collecting and storing precise location data, sharing it with friends or other users, and even gathering additional information like profile details and browsing history from other web applications. As a result, users face multiple risks, including data breaches that expose their information and the potential for physical harm if their location is constantly visible to criminals.

[4]Source not included in the REA but added into the narrative for context.

DISCUSSION AND CONCLUSION

This article highlights that conflict de-escalation in online communities, especially those focused on crime, safety, and neighbourhood vigilance, requires deliberate, context-sensitive intervention strategies. Unlike face-to-face encounters, where nonverbal cues, vocal tone, and social norms often assist in signalling empathy, backing down, or humour, digital interactions frequently lack this nuance. Misinterpretations, emotional overreactions, and performative posturing in front of a virtual audience can all heighten the potential for escalation. This is particularly true in crime-related digital spaces, where fear, moral outrage, and uncertainty can amplify tension. As such, moderators are not merely administrators but facilitators of digital public order who must intervene with both skill and ethical awareness.

The evidence provided through this REA underscores three foundational principles: neutrality, timeliness, and discretion. Neutrality ensures that moderators apply rules fairly and avoid appearing biased, which is critical to maintaining trust among users. Timely intervention prevents disputes from escalating by addressing issues before they intensify or spread. Discretion involves resolving conflicts through private channels when appropriate, reducing the risk of public embarrassment or defensiveness. Intervening early, before conflict becomes entrenched, can prevent hostile group dynamics from forming and reduce the emotional temperature of discussions. Private messaging is a recommended practice for addressing disputes in a less confrontational way. It removes the visibility of public shaming and reduces the potential for users to become defensive in front of an audience. While moderator neutrality, not taking sides or issuing value judgments during disputes, is essential for building trust and demonstrating procedural fairness. This does not mean ignoring harmful behaviour; rather, it involves applying rules consistently and being transparent about decision-making criteria. Importantly, moderators must be perceived as legitimate and impartial actors. Legitimacy in online governance, much like in physical communities, depends on the fair application of rules, transparent processes, and the opportunity for members to voice concerns or appeal decisions. When moderators are seen as biased or punitive, their actions may inadvertently reinforce division or lead to disengagement by marginalized group members. This is especially important in communities where cultural, linguistic, or political diversity exists, as the same behaviours may be interpreted differently depending on context.

Digital tools such as *ConsiderIt* and *TruthMapping* offer emerging solutions that support conflict resolution by encouraging structured, deliberative dialogue. These platforms allow users to express their viewpoints, consider opposing arguments, and reflect before reacting. While not yet widely adopted in mainstream community platforms like Facebook, they represent a shift toward digital deliberation – the idea that online engagement should mirror democratic norms of reasoned debate, not merely unmoderated expression. Integrating such tools into group moderation protocols, either directly or through adapted features, could enhance

the quality of online interactions and reduce impulsive, inflammatory posts.

Another key element identified in the literature is the value of post-conflict reflection and learning. After disputes are resolved, moderators and community administrators should conduct brief debriefs – either individually or as teams – to review the source of the conflict, evaluate whether the intervention was effective, and consider if changes to community rules or moderation protocols are necessary. This practice strengthens institutional memory, builds team competence, and enables the development of more inclusive and flexible moderation strategies over time. Platforms could support this by offering templates or guidance for conducting conflict audits. While automation and algorithmic moderation tools (such as flagging systems, profanity filters, and sentiment analysis) can assist in identifying problematic content, this review makes clear that automated systems are not yet capable of replacing human judgment. AI cannot discern sarcasm, cultural nuance, power imbalances, or historical tensions between users. Relying too heavily on automated detection can lead to over-enforcement in some communities and under-enforcement in others, particularly where marginalized groups use language differently or are disproportionately flagged. Human moderators bring critical interpretive skills to these situations: they can assess intention, offer explanation, and apply discretion, qualities essential for de-escalation.

This also raises important ethical considerations. The use of surveillance-based tools and predictive analytics to monitor potential conflict – particularly when tied to public safety and law enforcement – risks undermining user trust, especially if these systems are opaque or biased. While early warning systems may have a role to play, they must be accompanied by transparent governance frameworks, accountability measures, and clear limits on data use. Any safety strategy that relies on extensive surveillance must balance public protection with respect for privacy, civil liberties, and community self-governance.

The findings presented in this article point to several actionable recommendations for practice. First, it is important to invest in conflict de-escalation training for moderators. Moderators require structured training in mediation, neutral communication, cultural sensitivity, and trauma-informed responses. Conflict de-escalation should be viewed as a core competency, not an informal skill. Second, it is important to encourage private resolution mechanisms. Platforms and group administrators should enable or encourage private messaging options during disputes, paired with clear guidance on when and how to use them constructively. Third, implication for practice yielded from the REA is the significance of support-structured dialogue tools and reflective practice. Thus, the integration of features such as structured comment threads, argument mapping, or "cooling-off" delays for inflammatory posts can reduce impulsive responses and create space for reflection. Fourth is the promotion of transparency and procedural fairness, in that group rules, enforcement actions, and moderation procedures should be visible and consistently applied. Community members should be informed of why a post or comment was removed and given the opportunity to appeal. Fifth is the notion that ethical standards need to be embedded into platform and policing partnerships overseeing such work. Any involvement by law enforcement in digital communities should be guided by principles of proportionality, minimal intervention, and respect for user autonomy. This is especially important in marginalized communities that may have historical distrust of policing institutions. Finally, the sixth implication for practice centres on developing escalation protocols that include referral options. Therefore, where conflict involves threats, hate speech, or signs of mental health distress, moderators should be able to refer users to appropriate support services or escalate issues to platform safety teams or law enforcement where necessary.

In conclusion, de-escalating conflict in online communities is not a peripheral task, it is central to the health, trust, and cohesion of digital spaces. As online community groups increasingly serve as de facto forums for local safety, crime discussion, and civic engagement, the ability to manage disagreement respectfully and constructively becomes even more critical. Moderators, platform providers, and police must collaborate to ensure that digital spaces do not simply mirror the divisions of the offline world, but rather offer models of inclusive, respectful dialogue. This article affirms that effective de-escalation relies on a balance between procedural fairness, technological support, and human judgment. Community leaders must be equipped with both the tools and the training to manage disputes ethically, while platforms must ensure they provide environments conducive to dialogue, not division. Ultimately, a relational approach to moderation, grounded in trust, transparency, and shared norms, will be most effective in cultivating digital communities that are not only safe, but also socially resilient and democratically vibrant.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the support and assistance from the Queensland Police Service in undertaking this research. The views expressed in this publication are not necessarily those of the Queensland Police Service and any errors of omission or commission are the responsibility of the authors.

FUNDING

Queensland Police, DICE Prevention Project.

CONFLICT OF INTEREST DISCLOSURES

The authors have no conflicts of interest to declare.

AUTHOR AFFILIATIONS

*Faculty of Design & Society, University of Technology Sydney, Sydney, Australia; †Queensland Police Service, Brisbane, Queensland, Australia.

REFERENCES

Denotes publication used in REA presented in this article based on the research questions of 4, 5, 6, and 8.

*Alrayes, F., & Alia, A. (2014). No place to hide: A study of privacy concerns due to location sharing on geo-social networks. *International Journal on Advances in Security, 7*, 62–75.

Australian Institute of Health and Welfare (AIHW). (2024). *Family, domestic and sexual violence*. Retrieved February 12, 2025 from: https://www.aihw.gov.au/family-domestic-and-sexual-violence/types-of-violence/stalking-surveillance.

*Bailey, J. (2015). A perfect storm: How the online environment, social norms and law shape girls' lives. In J. Bailey & V. Steeves (Eds.), *eGirls, eCitizens: Putting technology, theory and policy into dialogue*

with girls' and young women's voices (pp. 21–54). University of Ottawa Press. https://doi.org/10.1353/book.40672

*Baur, C., & Prue, C. (2014). The CDC clear communication index is a new evidence-based tool to prepare and review health information. *Health Promotion Practice, 15*(5), 629–637. https://doi.org/10.1177/1524839914538969

*Berger, J. (2013). Beyond viral: Interpersonal communication in the internet age. *Psychological Inquiry, 24*(4), 293–296. https://doi.org/10.1080/1047840X.2013.842203

Berger, M., & Sklansky, D. A. (2023). Crime, community, and the shadow of the virtual. *University of Illinois Law Review, 2023*(5), 1607–1638.

*Bharati, T., Jetter, M., & Malik, M. N. (2022). *Types of communications technology and civil conflict*. I Z A - Institute of Labor Economics. https://www.jstor.org/stable/resrep64747

Bikku, T., Biyyapu, N. S., Sekhar, J. C., Kumar, M. K., Nokerov, S. M., & Pratap, V. K. (2024). The social network dilemma: Safeguarding privacy and security in an online community. *International Journal of Safety & Security Engineering*, *14*(1), 125–133. https://doi.org/10.18280/ijsse.140112

*Brownson, R. C., Eyler, A. A., Harris, J. K., Moore, J. B., & Tabak, R. G. (2018). Getting the word out: New approaches for disseminating public health science. *Journal of Public Health Management and Practice*, *24*(2), 102–111. https://doi.org/10.1097/PHH.0000000000000673

*Byman, D. (2017). How to hunt a lone wolf: Countering terrorists who act on their own. *Foreign Affairs, 96*, 96–105.

*Chan, A. K., Nickson, C. P., Rudolph, J. W., Lee, A., & Joynt, G. M. (2020). Social media for rapid knowledge dissemination: Early experience from the COVID-19 pandemic. *Anaesthesia, 75*(12), 1579–1582. https://doi.org/10.1111/anae.15057

*Cheong, N., Johns, A., & Byron, P. (2023). Queering the 'resourcing' of LGBTQ+ young people in the Asia Pacific. *Information, Communication & Society*, *26*(12), 2439–2456. https://doi.org/10.1080/1369118X.2023.2249970

*Cohen, K. (2013). *Who will be a lone wolf terrorist? Mechanisms of self-radicalisation and the possibility of detecting lone offender threats on the Internet*. Swedish Defence Research Agency. https://www.foi.se/rest-api/report/foi-r--3531--se

*Conway, M., Scrivens, R., & Macnair, L. (2019). *Right-wing extremists' persistent online presence: History and contemporary trends*. International Centre for Counter-Terrorism. https://www.jstor.org/stable/resrep19623

*Cubellis, M. A., Evans, D. N., & Fera, A. G. (2019). Sex offender stigma: An exploration of vigilantism against sex offenders. *Deviant Behavior, 40*(2), 225–239. https://doi.org/10.1080/01639625.2017.1420459

*De Bruyn, P. C. (2021). Evidence to explain violent extremist communication: A systematic review of individual-level empirical studies. *Perspectives on Terrorism*, *15*(4), 76–110. https://www.jstor.org/stable/27044237

*Dineva, D., & Breitsohl, J. (2022). Managing trolling in online communities: An organizational perspective. *Internet Research*, *32*(1), 292–311. https://doi.org/10.1108/INTR-08-2020-0462

*Dineva, D., & Daunt, K. L. (2023). Reframing online brand community management: Consumer conflicts, their consequences and moderation. *European Journal of Marketing*, *57*(10), 2653–2682. https://doi.org/10.1108/EJM-03-2022-0227

*Do, R., Kim, S., Lim, Y. B., Kim, S. J., Kwon, H., Kim, J. M., Lee, S., & Kim, B. N. (2023). Korean adolescents' coping strategies on self-harm, ADHD, insomnia during COVID-19: Text mining of social media big data. *Frontiers in Psychiatry, 14*, 1192123. https://doi.org/10.3389/fpsyt.2023.1192123

*Edelmann, N. (2013). Reviewing the definitions of "lurkers" and some implications for online research. *Cyberpsychology, Behavior, and Social Networking, 16*(9), 645–649. https://doi.org/10.1089/cyber.2012.0362

*Elsaesser, C. M., Patton, D. U., Kelley, A., Santiago, J., & Clarke, A. (2019). Avoiding fights on social media: Strategies youth leverage to navigate conflict in a digital era. *Journal of Community Psychology*, *49*(3), 806–821. https://doi.org/10.1002/jcop.22363

Evans, J. (2024). Laws to regulate misinformation online abandoned. *ABC News*. https://www.abc.net.au/news/2024-11-24/laws-to-regulate-misinformation-online-abandoned/104640488

Forestal, J. (2022). *Designing for democracy: How to build community in digital environments*. Oxford University Press. https://doi.org/10.1093/oso/9780197568750.001.0001

*Fourie, I., Agarwal, N. K., Sonnenwald, D. H., Julien, H., Rorissa, A., & Detlor, B. (2022). Everyday information behavior of marginalized communities in the global South: Informal transportation as example. *Proceedings of the Association for Information Science and Technology, 59*(1), 565–569. https://doi.org/10.1002/pra2.628

Frieden T. R. (2013). Six components necessary for effective public health program implementation. *American Journal of Public Health, 104*, 17–22.

*Ghouri, A. M., Akhtar, P., Vachkova, M., Shahbaz, M., Tiwari, A. K., & Palihawadana, D. (2020). Emancipatory ethical social media campaigns: Fostering relationship harmony and peace. *Journal of Business Ethics*, *164*(2), 287–300. https://doi.org/10.1007/s10551-019-04279-5

*Grimmelikhuijsen, S. G., & Meijer, A. J. (2015). Does twitter increase perceived police legitimacy? *Public Administration Review, 75*(4), 598–607. https://doi.org/10.1111/puar.12378

Hattingh, M. J. (2015, September). The use of Facebook by a Community Policing Forum to combat crime. In *Proceedings of the 2015 Annual Research Conference on South African Institute of Computer Scientists and Information Technologists* (pp. 1–10). ACM. https://doi.org/10.1145/2815782.2815811

*Hembroff, G. C., Boyle, D., & Wagner, T. V. (2020). The design of a holistic mHealth community library model and its impact on Empowering Rural America. In *Proceedings of the 8th EAI International Conference on Wireless Mobile Communication and Healthcare, MobiHealth 2019, Dublin, Ireland, November 14–15, 2019* (pp. 97–111). Springer International Publishing. https://doi.org/10.1007/978-3-030-49289-2_8

*Hernandez-Bocanegra, D. C., & Ziegler, J. (2021). ConvEx-DS: A dataset for conversational explanations in recommender systems. In *Interfaces and Human Decision Making for Recommender Systems 2021: Proceedings of the 8th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems* (pp. 3–20). CEUR Workshop Proceedings (CEUR-WS.org).

*Hirblinger, A. T. (2020). The strategic purposes of digital inclusion. In *Digital inclusion in mediated peace processes: How technology can enhance participation* (pp. 19–33). US Institute of Peace. http://www.jstor.org/stable/resrep26644.8

*Ho, A. T.-K., & Cho, W. (2017). Government communication effectiveness and satisfaction with police performance: A large-scale survey study. *Public Administration Review*, *77*(2), 228–239. https://doi.org/10.1111/puar.12563

*International Crisis Group. (2020). Political failures and escalating conflict. In *An Avoidable War: Politics and Armed Conflict in Myanmar's Rakhine State* (pp. 6–16). International Crisis Group. http://www.jstor.org/stable/resrep31467.6

*James, B., & Wilfred, W. W. (2024). The role of language in social media during the COVID-19 Pandemic. In *Public Health Communication Challenges to Minority and Indigenous Communities* (pp. 60–75). IGI Global.

*Jhaver, S., Ghoshal, S., Bruckman, A., & Gilbert, E. (2018). Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction, 25*(2), 1–33. https://doi.org/10.1145/3185593

Kleinig, J., Mameli, P., Miller, S., Salane, D., & Schwartz, A. (2011). The underlying values and their alignment. In *Security and privacy: Global standards for ethical identity management in contemporary liberal democratic states* (pp. 151–224). ANU Press. http://www.jstor.org/stable/j.ctt24h8h5.13

*Kriplean, T., Morgan, J., Freelon, D., Borning, A., & Bennett, L. (2013). Supporting reflective public thought with considerit. In *Proceedings of the ACM Conference on Computer Support Cooperative Work* (pp. 265–274). ACM.

*Kuan-Ming, C. (2024). Bystander behaviour in different cyberbullying situations: Role of cyberbullying awareness, moral disengagement, and victim behaviour. *Bulletin of Educational Psychology, 53*(3), 491–512.

*Lagmay, A. M. F. A., Racoma, B. A., Aracan, K. A., Alconis-Ayco, J., & Saddi, I. L. (2017). Disseminating near-real-time hazards information and flood maps in the Philippines through Web-GIS. *Journal of Environmental Sciences, 59*, 13–23. https://doi.org/10.1016/j.jes.2017.03.014

*Lane, J., & Stuart, F. (2022). How social media use mitigates urban violence: Communication visibility and third-party intervention processes in digital urban contexts. *Qualitative Sociology, 45*(3), 457–475. https://doi.org/10.1007/s11133-022-09510-w

*Lee, J. R., Holt, T. J., & Warren, I. (2021). Big data, cyber security and liberty. In B. A. Arrigo & B. G. Sellers (Eds.), *The pre-crime society: Crime, culture and control in the ultramodern age* (1st ed., pp. 409–432). Bristol University Press. https://doi.org/10.2307/j.ctv1rnpjdp.23

*Leman-Langlois, S., Campana, A., & Tanner, S. (2024). Policing the far right. In *The Great Right North: Inside Far-Right Activism in Canada* (Vol. 267, pp. 194–228). McGill-Queen's University Press. https://doi.org/10.2307/jj.20829378.10

*Letouzé, E., Meier, P., & Vinck, P. (2013). Big data for conflict prevention: New oil and old fires. In F. Mancini (Ed.), *New Technology and the Prevention of Violence and Conflict* (pp. 4–27). International Peace Institute.

*Lowry, P. B., Zhang, J., Wang, C., & Siponen, M. (2016). Why do adults engage in cyberbullying on social media? An integration of online disinhibition and deindividuation effects with the social structure and social learning model. *Information Systems Research, 27*(4), 962–986. https://doi.org/10.1287/isre.2016.0671

*Lowry, P. B., Moody, G. D., & Chatterjee, S. (2017). Using IT design to prevent cyberbullying. *Journal of Management Information Systems, 34*(3), 863–901. https://doi.org/10.1080/07421222.2017.1373012

*MacCarthy, M. (2023). Privacy rules for digital industries. In *Regulating digital industries: How public oversight can encourage competition, protect privacy, and ensure free speech* (pp. 171–230). Brookings Institution Press.

*Marret, M. J., & Choo, W. Y. (2017). Factors associated with online victimisation among Malaysian adolescents who use social networking sites: A cross-sectional study. *Br Med J Open, 7*(6), e014959. https://doi.org/10.1136/bmjopen-2016-014959

*Martin, L. J., Turnquist, A., Groot, B., Huang, S. Y., Kok, E., Thoma, B., & van Merriënboer, J. J. (2019). Exploring the role of infographics for summarizing medical literature. *Health Professions Education, 5*(1), 48–57. https://doi.org/10.1016/j.hpe.2018.03.005

*Marwick, A., & Boyd, D. (2014). 'It's just drama': Teen perspectives on conflict and aggression in a networked era. *Journal of Youth Studies, 17*(9), 1187–1204. https://doi.org/10.1080/13676261.2014.901493

*McMahon, M. (n.d.). Should suspects in criminal cases be publicly named? *Deakin University*. https://this.deakin.edu.au/society/should-suspects-in-criminal-cases-be-publicly-named

McMahon, L., Kleinman, Z., & Subramanian, C. (2025, January). Facebook and Instagram get rid of fact checkers. *BBC News*. https://www.bbc.com/news/articles/cly74mpy8klo

*Miller, K. C., & Lewis, S. C. (2023). Journalistic visibility as celebrity and its consequences for harassment. *Digital Journalism, 11*(10), 1886–1905. https://doi.org/10.1080/21670811.2022.2136729

*Milofsky, A., Sany, J., Lancaster, I., & Krentel, J. (2017). *Conflict management training for peacekeepers: Assessment and recommendations*. US Institute of Peace.

*Moore, C. W. (2014). *The mediation process: Practical strategies for resolving conflict*. John Wiley & Sons.

*Morah, D. N., & Oladokun, O. (2023). Cross-regional analysis of terrorism reporting and dynamics of ethnic relations in Nigeria. In *Research Anthology on Modern Violence and Its Impact on Society* (pp. 346–362). IGI Global.

*Nabiullina, V. R. (2021). Public dissemination of knowingly false information: Criminal and criminological aspects. In *Proceedings of the VII International Scientific-Practical Conference "Criminal Law and Operative Search Activities: Problems of Legislation, Science and Practice"* (pp. 70–74). https://doi.org/10.5220/0010629000003152

*Ng, F. K., Wallace, S., Coe, B., Owen, A., Lynch, J., Bonvento, B., & McGrath, B. A. (2020). From smartphone to bed-side: Exploring the use of social media to disseminate recommendations from the National Tracheostomy Safety Project to front-line clinical staff. *Anaesthesia, 75*(2), 227–233. https://doi.org/10.1111/anae.14747

Parker, M., & Dodge, M. (2024). The new neighborhood watch: An exploratory study of the nextdoor app and crime narratives. *International Journal of Criminology and Sociology, 13*, 43–54. https://doi.org/10.6000/1929-4409.2024.13.04

*Pauwels, E. (2020). *Artificial intelligence and data capture technologies in violence and conflict prevention: Opportunities and challenges for the international community*. Global Center on Cooperative Security. https://www.jstor.org/stable/resrep27551

*Peter, J., & Valkenburg, P. M. (2013). The effects of internet communication on adolescents' psychosocial development: Anassessment of risks and opportunities. In A. N. Valdivia (Ed.), *The international encyclopedia of media studies*. Wiley-Blackwell.

*Qadir, S., Niser, A., Caddle, X. V., Alsoubai, A., Park, J. K., & Wisniewski, P. J. (2024, May). Towards a safer digital future: Exploring stakeholder perspectives on creating a sustainable youth online safety community. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1–10). https://doi.org/10.1145/3613905.3651019

*Rodilosso, E. (2024). Filter bubbles and the unfeeling: How AI for social media can foster extremism and polarization. *Philosophy & Technology, 37*(2), 71. https://doi.org/10.1007/s13347-024-00758-4

*Rosenstiel, T., Sonderman, J., Loker, K., Ivancin, M., & Kjarval, N. (2015). *Twitter and the news: How people use the social network to learn about the world*. American Press Institute. Retrieved from https://americanpressinstitute.org/wp-content/uploads/2015/09/Twitter-and-News-How-people-use-Twitter-to-get-news-American-Press-Institute.pdf

*Roystonn, K., AshaRani, P. V., Devi, F., Wang, P., Zhang, Y., Jeyagurunathan, A., & Subramaniam, M. (2023). Exploring views and experiences of the general public's adoption of digital technologies for healthy lifestyle in Singapore: A qualitative study. *Frontiers in Public Health, 11*, 1227146. https://doi.org/10.3389/fpubh.2023.1227146

*Sachdeva, N., & Kumaraguru, P. (2015). Online social media and police in India: Behavior, perceptions, challenges. arXiv preprint arXiv:1403.2042.

*Saleem, H. M., Dillon, K. P., Benesch, S., & Ruths, D. (2017). A web of hate: Tackling hateful speech in online social spaces. arXiv preprint arXiv:1709.10159.

*Schwartz, H. L., Ramchand, R., Barnes-Proby, D., Grant, S., Jackson, B. A., Leuschner, K. J., Matsuda, M., & Saunders, J. (2016). Using innovative technology to enhance school safety in practice. In *The role of technology in improving K-12 school safety* (pp. 35–54). RAND Corporation.

*Sigsworth, R. (2019). *#SpeakUp: Using social media to promote police accountability in Kenya, Tanzania and Uganda*. Institute for Security Studies. https://www.jstor.org/stable/resrep62833

*Simon, T., Goldberg, A., & Adini, B. (2015). Socializing in emergencies: A review of the use of social media in emergency situations. *International Journal of Information Management, 35*(5), 609–619. https://doi.org/10.1016/j.ijinfomgt.2015.07.001

*Singh, J. P. (2013). Information technologies, meta-power, and transformations in global politics. *International Studies Review, 15*(1), 5–29. https://doi.org/10.1111/misr.12025

*Sullivan, C., Staib, A., McNeil, K., Rosengren, D., & Johnson, I. (2019). Queensland digital health clinical charter: A clinical consensus statement on priorities for digital health in hospitals. *Australian Health Review, 44*(5), 661–665. https://doi.org/10.1071/AH19067

*Sun, N., Rau, P., & Ma, L. (2014). Understanding lurkers online communities: A literature review. *Computers in Human Behavior, 38*, 110–117. https://doi.org/10.1016/j.chb.2014.05.022

*Thoma, B., Murray, H., Huang, S. Y. M., Milne, W. K., Martin, L. J., Bond, C. M., Mohindra R., Chin A, Yeh C. H., Sanderson W. B., & Chan, T. M. (2018). The impact of social media promotion with infographics and podcasts on research dissemination and readership. *Canadian Journal of Emergency Medicine, 20*(2), 300–306. https://doi.org/10.1017/cem.2017.394

Tierney, T. (2013). *The public space of social media: Connected cultures of the network society*. Routledge.

*Tim, Y. N., Pan, S. L., Bahri, S., & Fauzi, A. (2014). Social media as boundary objects: A case of digitalized civic engagement in Malaysia. In *Proceedings of the 25th Australasian Conference on Information Systems, 8th–10th December, Auckland, New Zealand*. ACIS. https://openrepository.aut.ac.nz/collections/5b7854aa-901e-4c5c-a457-7928503cb23f

Trottier, D. (2020). Denunciation and doxing: Towards a conceptual model of digital vigilantism. *Global Crime, 21*(3–4), 196–212. https://doi.org/10.1080/17440572.2019.1591952

*Türker, G., & Gök, A. (2024). Video Games and Radical Movements: "EIN PROZENT" and "HEIMAT DEFENDER." *Journal of Strategic Security, 17*(2), 89–125. https://doi.org/10.5038/1944-0472.17.2.2218

Tyagi, A. K., Naithani, K., & Tiwari, S. (2024). Security and possible threats in today's online social networking platforms. In S. K. Rathi, B. Keswani, R. K. Saxena, S. K. Kapoor, S. Gupta, & R. Rawat (Eds.), *Online social networks in business frameworks* (pp. 159–199). https://doi.org/10.1002/9781394231126.ch8

*Van Mierlo, T. (2014). The 1% rule in four digital health social networks: An observational study. *Journal of Medical Internet Research, 16*(2), e33. https://doi.org/10.2196/jmir.2966

*Zhu, H., Wu, H., Cao, J., Gang, F., & Li, H. (2018). Information dissemination model for social media with constant updates. *Physica A: Statistical Mechanics and its Applications, 502*, 469–482. https://doi.org/10.1016/j.physa.2018.02.142