Toda Peace Institute

Policy Brief No. 34 March 2019

Counteracting Hate and Dangerous Speech Online:
Strategies and Considerations

Rachel Brown and Laura Livingston

Abstract

This policy brief will examine the various factors that enable online hate speech to resonate, spread, and drive offline action. After briefly reviewing the features of social media that enable hate speech to spread online, we will explore tools for designing interventions to respond to this content. As part of this, we will consider the broader online and offline context impacting this speech, and review approaches to identifying, understanding, and engaging online audiences. Drawing from multidisciplinary research insights, the discussion will then address considerations for developing messaging strategies and content. The review concludes with a brief discussion of the importance of assessing and mitigating risk. Overall, this brief will position readers to be able to develop their own strategies for responding to online hate and dangerous speech in their context.

Social Media, Speech, and Violence

1. Communication is a powerful tool that can be used to either unite or divide. We have seen it weaponised to drive polarisation, violence, even genocide. This type of weaponised communication follows clear patterns that transcend history and geography: pitting an existentially good, righteous, and vulnerable "us," against a dangerous, guilty, and even sub-human "other" or "them." These communications create strong normative pressures for people to participate in or condone violence, and justify violence through narratives of self-defense, revenge, and protection.[i]

2. This policy brief will specifically focus on how to understand and respond to online communications that increase the risk of inter-group and group-targeted violence. Indeed, recent research has shown that exposure to hate speech desensitises individuals to this type of verbal aggression and increases their prejudice towards those targeted in the speech,[ii] suggesting that even encountering hate speech online undermines peaceful intergroup relations. This is especially relevant as new or increasingly popular online platforms have enabled broader public participation in both promoting and combatting this communication.

3. There are various terms and definitions for the communication of concern - colloquially we know it as "hate speech," while tribunals have labeled communication that contributes to violence "incitement."[iii] "Dangerous speech" has been defined as "any form of expression (e.g. speech, text, or images) that can increase the risk that its audience will condone or commit violence against members of another group."[iv] Beyond the speech's content, this concept focuses on the speech's ability to influence its audience by examining the audience itself, the surrounding social and political context, how it is disseminated, and the speaker's credibility and influence.

What is Unique about Social Media?

4. Throughout history, as new channels for communication have emerged, they have tapped into and interacted with the surrounding context and dynamics, at times fomenting and amplifying intergroup tensions, protest movements,[v] and violence - even genocide.[vi] Social media, as a new communication platform, inevitably impacts the information ecosystem and broader offline context. In so doing, these platforms create both challenges and opportunities for responding to harmful communications online. Below are some of the ways that social media platforms change the nature of communication that contributes towards violence in several ways.

5. Social media removes traditional media gatekeepers and democratises content creation, enabling communication to reach farther and faster. It has the power to amplify voices advocating and demonstrating peace and unity as desirable and expected behaviors. Platform users may also build and embrace new online communities that geographic limitations would otherwise preclude – communities that transcend location and traditional conflict or dividing lines.[vii]

6. These same features also allow fringe movements to develop, grow, and even seem mainstream. For those already harbouring discriminatory views, the internet provides limitless information to confirm hateful notions and connect with like-minded individuals. In doing so, they can enter new normative environments, breaking free of constraints that exist offline. Consider how Dylann Roof's search for "black on white crime" connected him to scores of misleading statistics and narratives on the issue, as well as people promoting those viewpoints, making it seem all the more prevalent a problem.[viii]

7. In privileging content that draws the longest or most extensive engagement from users – typically content that appeals to negative primal emotions (fear, anger, disgust) – platform algorithms can end up promoting fear and division. These posts often tap into divisive identities of "us" and "them" – driving our newsfeeds to create new, or intensify

existing, intergroup divisions and distrust. Because algorithms are programmed to give us more of what we like (the viewpoints, posts, or pages similar to those with which we have previously engaged), we are fed additional information that further reinforces these views.[ix]

8. Bots and other automated accounts programmed to further promote polarising content then amplify this content's reach, driving further division. In their contribution shaping the online space and the types of information shared, these bots also shape our perception of what normal behaviour (including expressed beliefs) is for our online communities.[x]

What can we do about dangerous online speech? Below, we emphasise the importance of understanding the broader context and the specific audience one hopes to reach; we then outline key strategies and considerations for developing interventions.

Analyze the Context to Inform Action

While online spaces certainly contribute to spreading and sustaining hateful content, online rhetoric does not exist in a vacuum. Online content taps into, reinforces, and even supercharges salient narratives, long-standing grievances, intergroup divisions and conflicts deeply rooted within a given context.[xi] This means that offline information sources and credible messengers – whether local leaders, information spreaders, or other influential community members – also impact online information flows. Moreover, content may jump across multiple offline and online mediums: consider how a print news story gets posted onto Facebook, where someone screen-grabs the headline and sends it to a closed Messenger or WhatsApp group, whose members then discuss it offline.

As a contemporary example, consider the outsized role that influential members of South Sudan's diaspora have played in circulating inciting Facebook posts. These posts tap into longstanding tensions and a history of violence between the Dinka and Nuer tribes in South Sudan and thus resonate with South Sudan's in-country population. In-country South Sudanese then spread this content both online and by word of mouth, enabling messages from outside the country to reach and influence the in-country offline population.[xii]

Any intervention in response to online hate speech must take this broader context into account – both the influence of our human psychology in driving us to believe in or spread this content, and the interplay between different online and offline mediums.

The context analysis below reviews the web of factors that allow the concerned online content to spread and resonate. Collectively, this analysis allows us to understand the web of dynamics that impact whether, how, and with whom online hate speech resonates and drives to offline action. Though each component is described individually, the components overlap and interact with one another. After exploring each factor, we discuss their collective application using the contemporary example of Sri Lanka.

9. **Information ecosystem:** In seeking to address hate speech on a particular platform, consider not only how the platform operates, but also how it fits into the target audience's larger information ecosystem. This includes considering: whether the platform is something that individuals rely on as their primary source of information; whether

and how they share the information they receive on the platform; how the platform interacts with other communication mediums, including other social media sites or messaging apps; and, whether individuals find the platform a credible or reliable source of information.

Such an analysis can create a greater understanding of how hate speech on a particular platform might spread. This can better inform intervention strategies, specifically whether an intervention should target only the concerned platform or instead work across multiple mediums. For instance, if someone is initially encountering hateful rhetoric on Facebook and later sharing it on Twitter, merely responding to hateful tweets will not sufficiently address the spread of the content.

10. **History:** Consider whether the hateful content is grounded in a longstanding history of tension, conflict, or other grievances between groups. How is this history told and understood by different groups? Is there an alternative history of cooperation between the two groups that your intervention can instead emphasise?

11. **Narratives**: Narratives are collective stories that frame individuals' understanding of the world around them and their place within it. Narratives rely on underlying proof points (events, statistics, news stories) as evidence or support for their particular story. For example, a story of an individual committing a crime could be used as a proof point for the narrative that an entire group of people – seen as represented by that individual – is dangerous. Efforts to respond to online hate speech should consider not only the narratives used to spread or reinforce hatred and division, but also those emphasising unity and inclusion within a given context.

12. **Identities:** It is important to map out and understand the identities that are being activated for conflict, as well as those that may provide cross-cutting ties that could be leveraged to unify people across conflict lines (e.g., mothers, citizens of a town, music or sports fans, etc.).

13. **Relevant actors:** It is key to understand relevant actors – both those promoting hate and dangerous speech, and those pushing back. These relevant actors might include influential social media personalities, previously offline voices that now have a bigger platform online, or even one's online peers with whom geographic constraints previously precluded relationships (e.g., diaspora groups). In addition to addressing those promoting negative speech, an intervention can seek to support (and learn from) those already pushing back.

14. **Contemporary example: Social media and religious tensions in Sri Lanka:** On multiple occasions in Sri Lanka, offline altercations between Buddhist and Muslim Sri Lankans were recorded and uploaded to Facebook, fueling anti-Muslim rumours, deadly protests, and revenge attacks throughout Sri Lanka. The videos tapped into Buddhists' historical narrative of being under threat from a minority population, including a decades-long civil war against the country's Tamil minority (historical context, narrative context). These rumours depicted Muslims as plotting to sterilize and ultimately wipe out the Sinhalese (identity context). Local community leaders and extremist

voices seized on both incidents as proof points for the Muslim plot, using their credibility and large online and offline networks to spread the content and urge the Sinhalese to take up arms to defend themselves (narrative context, actors' context). In addition to speaking with community members in offline forums, leaders uploaded Facebook videos that later shifted to private WhatsApp groups to call for the Sinhalese to take up arms to reclaim their country (information ecosystem).[xiii]

Choose an Audience and Set Goals

15. **The audience's perspective comes first**: As you develop your overall strategy, consider the audience's perspective – including their logic and emotions – in developing content. Once you've chosen an audience to focus on, learn as much as you can about them. For instance, you might seek to determine: Information about where and how they access information, and who they trust, and look to for guidance. This will help you better understand the world from the audience's perspective, enhancing the influence of your interventions. Consider how your message fits into the broader picture of your target audience's life – their existing beliefs, values, social networks, and their related emotions and experiences.

    Before learning deeply about an audience's perspective, it is important to consider which audiences you want to reach. One way to map out different audiences is by considering both attitudes towards peace and inclusion or towards hate and dangerous speech (which resonates with an audience?), and then by considering an audience's level of involvement in taking relevant actions (is an audience actually engaging in the relevant speech or behaviors).

16. **Increase new audiences with a positive attitude and low involvement**: These are people who may not like the surrounding hate/dangerous speech but may also not be taking any action to counter it or provide alternative messages. The goal for this group is that they take action – for example, by speaking up. This has the potential to shift the perceived norms in the online space, and in doing so counter one of the most important impacts of hate and dangerous speech: its ability to silence those who oppose it by increasing the perceived costs of speaking out.

17. **Increase new audiences with a positive attitude who are already taking action**: You can try to support these groups in remaining active by providing encouragement, elevating their voices to new audiences, or reaching out and asking them how to get involved.

18. **Reduce audiences with a negative attitude and high involvement:** These are people with an affinity for dangerous speech who are taking action. The goals for this group are two-fold: to reduce their engagement/involvement (a behavioural goal), or to shift their attitudes/feelings so that they become more neutral. Targeting this group is likely to be most effective if there are a few key influencers who can be reached through long-term and high-level engagement, or through initiatives that specifically deal with extreme groups.

19. **Reduce audiences with a negative attitudes and low involvement**: These are people with an affinity for dangerous speech who aren't taking action (maybe they are constrained by norms, don't have time, etc.). For this group, we want to prevent them from becoming more active and want to shift their attitudes and feelings so that they become less negative.

20. **Engage the disengaged:** There will always be people who could go either way: they don't have a strong positive or negative attitude, feeling, or affinity. Some may be highly engaged in general – they may be involved in local civic life, be big information producers, influential in their communities, or they may be less involved. Either way, you can seek to engage them positively, and to move them from a neutral attitude to a more positive attitude.

Consider Promising Strategies

Once you have thought about your audience, you can consider some promising approaches and related risks:

21. **The challenges of debating**: When we want to change someone's mind, it is intuitive to try to do so by arguing and debating. Unfortunately, this is often ineffective and may even backfire, causing people to cling to their existing beliefs even more strongly than before. The psychological mechanism behind this is *motivated reasoning,* whereby people often reject information that challenges their beliefs, while seeking out and accepting information that confirms their beliefs. This especially occurs with beliefs that challenge one's sense of self or group identification.[xiv] In practical terms, this means that interventions designed to correct or stop the spread of hateful content, if not done carefully, may achieve the opposite.

22. **Provide opportunities for people to change their own minds:** It is not easy to change our minds, admit we are wrong, or leave groups of which we are part. Doing these things often leaves us feeling vulnerable. However, there are ways to make these choices seem less threatening. Consider employing journey stories in your outreach, particularly those that show role models or like-minded individuals who questioned their past beliefs and behaviours, or who decided to stop remaining silent in the face of hateful content. You can also re-package information within an audience's existing narrative (i.e. rather than asking someone to dismiss a grievance, you can affirm that grievance but provide an alternate explanation that does not implicate a target group). Finally, you can connect new information to values (i.e., fairness, caring, independence) that matter to your audience.

23. **Build trust over time**: If you are going to challenge beliefs, it is important to first build credibility and trust as a messenger, or to partner with people your target audience trusts. Also, ensure any new information that you are providing comes from sources your audience finds credible. Finally, be patient and stay engaged – changing beliefs can be a long-term process and is unlikely to happen in the course of one conversation. Prepare to stay engaged over time and remember that how you treat those who disagree with you during the engagement can matter just as much as the discussion that takes place.

24. **Avoid shame:** Shaming someone into changing their mind is likely to backfire. Shame is the belief that one is inherently bad ("I am a bad person") and arises from things like name-calling. Shame is distinct from guilt, which focuses on an action ("I did something bad"). While a person can take positive action to address a mistake, feeling shame or humiliation can cause people to turn inward, become defensive, or dig-in to an existing belief.[xv]

25. **Use best practices when challenging falsehoods:** Consider the below insights from research on successful attempts to correct misinformation:[xvi]

    - Use a source that is not considered ideologically aligned with the content of the correction (one that, ideally, your audience already trusts);

    - Provide an alternative causal explanation of why the misperception occurred;

    - Avoid repeating the false claim itself, even if repeating it with a negation. This is because the more familiar we are with a piece of information, the more likely we are to believe it is true and even a negation increases familiarity. For instance, if the false claim is "John is a criminal," stating "John is *not* a criminal" may reinforce the core components of the statement ("John" and "criminal"), strengthening the John-criminal association that the statement intended to falsify.

    - If you must repeat a false claim, provide a warning first.

Note: It is particularly important to challenge narratives that portray the targeted group(s) as threatening (to a way of life, to physical security, values, etc.) or as guilty of violating core moral values, perpetrating violence, or other wrongs.

26. **Pay attention to tone**: Consider what kind of language is used and understood by the audience groups you aim to influence. Remember that speech extends beyond language itself – it can include emojis, images, charts, or other multimedia.

27. **Be aware of self-justification**: When people engage in negative behavior (i.e. sharing hateful content on Facebook), they tend to justify that behavior to avoid feeling negatively about themselves and their actions.[xvii] If you are interested in reaching people who have already taken negative actions, consider how you can enable them to change their behaviour without threatening their idea of themselves as good people.

28. **Media literacy**: At a more macro level, the internet requires a new media literacy – one that addresses the role of algorithms, filter bubbles, clickbait, automated accounts, and the monetary incentives underlying these factors. To increase our own agency in our online actions, this media literacy needs to provide us with greater self-awareness for how these factors interact with our human tendencies and biases. Interventions can seek to provide this type of self-awareness.

Create Positive Norms and Opportunities for Action

Our perception of our peers' beliefs or behaviours strongly influences our actions, even if those actions are contrary to our privately-held beliefs.[xviii] Online hate speech can project that hatred and prejudice are normal or expected towards a given group, or create the impression that, to be a part of the online community, we have to share similar content. Research even suggests that viewing others' comments on a particular post or topic may influence our reaction to the issue concerned.[xix] Even when deciding not to share hate speech, we may be more reluctant to push back on such comments as doing so (and violating perceived social norms) may lead to social ostracism, doxing, trolling, among other consequences.

By activating and elevating positive participation and new voices, you can decrease the power of hate and dangerous speech in your context. One way to do this is by activating audience members who have a positive or neutral attitude but who are disengaged.

Consider the following strategies for driving positive action:

29. **Avoid implying negative norms:** Interventions should avoid depicting hate or discrimination as normal or expected behaviour ("hatred toward group X is everywhere"), because it can have the unintended consequence of increasing acceptance of or participation in the negative behaviour or attitudes. When acknowledging negative behaviours or actions, be sure to also indicate that most people don't approve of the concerned actions or speech.

30. **Create positive social norms**: It is possible to create positive social norms by providing role models and content that show relevant peers taking action or speaking up. It can also be done through providing statistics or insights about a general group that highlight the prevalence of positive actions or attitudes. You can also focus on elevating stories about positive actions and speech especially from key role models and influential people.

31. **Make it easier to act:** Consider creating ladders of engagement to demonstrate small, incremental steps for people to become engaged (rather than asking someone to go from 0 to 100). A series of steps might include: liking a post, making a comment, sharing a post, creating a post, creating a page, organising an event, etc. Providing these concrete options is helpful when your audience is interested in acting but doesn't know what to do.

    Relatedly, consider providing spaces for collective action. It can be scary to act alone, especially when risks are involved (being called out, harassed, trolled, etc.). Consider creating spaces for individuals to act as a group to avoid the risk that they will be singled out, and to create a sense of belonging around taking positive action. For example, giving people the opportunity to join a group, participate in a campaign, or support and elevate one another's content.[xx]

Maintain, Create, and Reinforce Unifying and Cross-Cutting Identities

32. **Emphasise cross-cutting identities & build unifying identities:** We all have multiple social identities, and the identities that are (or are made) salient impact our behavior and attitudes towards particular groups. Are there ways to emphasise existing cross-cutting identities (those that transcend lines of conflict) that can bring people together across traditional dividing lines (e.g., using a sports team to unite fans across

religious lines or a religious identity to bring people together across ethnic lines)? Are there ways to build new identities that unify people? Consider the possible approaches below:

- Creating a space (e.g., a hashtag, Facebook, page) that brings people together around a shared identity that includes people from multiple backgrounds and across lines of division. This may be especially important if there is a recent growth in online spaces promoting hate or division and a seeming absence of those for finding a more peaceful or neutral community. Emphasise a shared identity in a comment responding to a particular hateful post, "As a mother…"
- Identify and emphasise shared aspirations;
- Use a brand to create a new unifying identity. A brand can be used to spark a movement or a set of behaviours and prevent people from feeling that they are acting alone; it can provide meaning, consistency, and show collective action. In high risk contexts in particular, a brand can also provide a way for people to speak out together and with some cover, especially if they may face risks of retaliation.

Challenge Narratives of Targeted Groups

Consider how you can change the narratives around targeted groups from how they are commonly stereotyped to a more positive or nuanced depiction. (While it is tempting to simply focus on generating empathy for groups, this alone may not lead to change action.)

33. **Challenge dehumanising stereotypes**: Dehumanisation is the idea that a group of individuals is not fully human – whether because they don't have secondary emotions seen as uniquely human (nostalgia, hope, disappointment), full cognitive ability, and/or are otherwise not evolved or civilised.[xxi] Groups that are dehumanised are often stereotyped as having low competence and low warmth (human emotions, compassion, and positive intention).[xxii] Efforts to humanise should consider the way in which a group is dehumanised and attempt to ensure that they are seen as having high levels of both warmth and competence.

34. **Challenge meta-perceptions (especially meta-dehumanisation):** In the context of intergroup relations, meta-perceptions relate to how we believe "they" perceive "us." For example, research has shown that when we believe another group dehumanises our group, we dehumanise their group in return ("reciprocal dehumanisation" or "retaliatory dehumanisation").[xxiii] In essence, content depicting an out-group as hating one's in-group may lead in-group members to hate that group in return. In engaging with such content online, consider how you can correct misperceptions about intergroup dehumanization and instead emphasise instances of intergroup cooperation.

    Provide content that challenges the belief that the concerned out-group dehumanizes one's in-group. This can be done through demonstrating inter-group friendships or relationships or even through stories, quotes, statements, or actions from the group that's being targeted.

35. **Consider empathy carefully:** When a group is being targeted, it is often tempting to generate empathy for the concerned group. However, while empathy can be influential, generating empathy does not always yield the desired behavior. For example, it's possible that empathy toward a targeted group yet can be outweighed by empathy for our own group. Such empathy can also lose out to other influences (i.e. a desire to protect our family or fit in with our peers). There may be barriers other than lack of empathy that need to be addressed in order to change behavior.

    Efforts to generate empathy for an out-group may also inadvertently lead to negative reactions,[xxiv] – or even to dehumanization. Inadvertent dehumanization may happen through portraying a targeted group of people as victims or as lacking agency, or through inadvertently evoking disgust or other strong negative emotions (i.e. showing someone next to a pile of garbage), or by depicting the targeted group as lacking complex emotions (nostalgia, hope, disappointment) or thoughts.

36. **Combine empathy and humanization**: This may be done through focusing on targeted groups' secondary (more complex human) emotions and decision-making processes – nostalgia, hope, disappointment, concern, through showing people in situations relatable to the audience, and through showing people demonstrating warmth and compassion.[xxv]

37. **Combine empathy and social norms:** Witnessing other group members showing empathy may lead individuals to themselves experience greater empathy and act more pro-socially toward out-group members. For example, a communication intervention might show an in-group member taking positive action toward a targeted out-group member.[xxvi]

38. **Use stories**: People are often more open to (and less defensive in the face of) stories, including stories concerning targeted out-groups. Consider creating stories to challenge existing narratives and opening new channels for people to share their own stories across divides.

Change the Conversation

39. **Framing**: Consider if you can reframe a conversation that targets a specific group to instead focus on the issue itself and the underlying grievance. This will allow you to address legitimate grievances and concerns while avoiding scapegoating a particular group as responsible for those issues. For instance, if your target audience is concerned about the economy and is blaming Group X for stealing their jobs, see if you can reframe the issue as a socioeconomic or political issue (rather than an ethnic one).

40. **Select a messenger (or messengers) carefully.** The messenger is just as important as the message. Especially in highly divided contexts, a relevant, credible, influential messenger will increase the likelihood that your audience will react positively to your message, while the wrong speaker (one lacking credibility (or even seen as mal-intentioned) among your audience) can discredit the message. Messengers can either directly deliver the content or they can encourage people to participate in the intervention. In selecting and recruiting messengers to participate in an intervention, here are some things to consider.

41. **Tap into existing social networks to find influencers** who have already built connections and audiences to activate them to spread positive messages. For example, consider messengers who have already build their own large audience, who play an important role in spreading information in general, or who are influential for a particular group (e.g., local faith leaders, elders, opinion leaders).

42. **Select speakers who can model desired behaviors and attitudes.** Positive norm setters can signal that it is more socially acceptable and safe to take positive action.

43. **Consider fictional speakers:** Sometimes it can be hard to find the ideal messenger. You can also consider building a fictional speaker (e.g., a character) or a non-human speaker (a brand) to spread messaging. In highly polarized environments where individuals are likely to be painted as belonging to one side or another, a fictional character can break beyond existing divisions while a brand can form a new identity with the ability to bring people together across lines of division.

44. **Consider using a surprise speaker**: People may be more likely to believe or pay attention to someone who they don't expect to share a specific message – either because it's unusual or because it seems against their interest. Is there a messenger who can serve as a surprise speaker to generate attention and increase the likelihood that people will consider the information? Or a group of speakers people don't expect to see together? For instance, consider using a person of a particular political party questioning the statistics cited to support that party's proposed policy.

45. **Make the audience a speaker**: Your strategy can directly engage your audience in becoming a speaker. Perhaps you want to empower audience members to share messages in support of peace or inclusion on Facebook or to retweet a specific image advocating the same.

Create a Strategy to Use Impactful Mediums

In addressing hate and dangerous speech on social media, it's still important to consider the way that social media interacts with other mediums, and to consider an approach that connects across mediums and platforms and allows your message to reach your audience. Here are some tips:

46. **Use an integrated strategy**: Audience members rely on different mediums (online and offline) for different purposes. While social media facilitates spreading messages to larger and broader audiences, messages that people hear directly from friends or other trusted messengers may also have critical impact. For example, rumours circulating offline may make their way into Facebook posts that are tweeted and later screen-grabbed and sent through closed WhatsApp or Messenger groups, allowing previously offline rumours to spread across multiple mediums. An intervention that solely focused on rumours on Facebook would miss how the information had spread (and perhaps evolved) over other mediums. With this in mind, strategies to counter negative online speech should consider which mediums are relevant and build interventions that span the multiple relevant online (and offline) platforms, and that encourage the spreading of positive speech both across and within platforms.

47. **Build on existing behaviors**: It is much easier to engage people in using a medium they already rely on than to drive them to use a new medium. If your audience members already heavily rely on Facebook or Instagram for spreading information, develop a strategy that can build on these existing habits.

48. **Consider breaking into echo chambers**: Platform algorithms are designed to feed us information that reinforces our existing worldview and belief system, ultimately reinforcing and intensifying existing biases. Consider how you can break into these echo chambers or filter bubbles to carefully introduce new information to which group members are otherwise not exposed.

Always Consider Risk

Perhaps most importantly, keep in mind that there is always the risk that a well-intentioned effort could backfire or cause unintended harm. For example, efforts to counter a fearmongering rumour can fuel it if they repeat the misinformation; attacks on people's identity can make them close ranks; calling attention to negative behavior in the wrong way can increase perceptions of a negative norm. As you think about taking action, be sure to first do no harm by always considering whether there are any risks: Is there a risk that your intervention could strengthen the impact of the very speech/communication you are hoping to counter (e.g., by raising its profile, making audiences more accepting, etc.)? Is there a risk that a particular action or programme could undermine your/your organization's ability to do work in the long-term (e.g., by damaging your credibility)? And could a particular action or programme pose risks to individuals and organizations involved (e.g., physical risks, legal risks, reputational risks, etc.).

The Authors

Rachel Brown is the Founder and Executive Director of Over Zero, the author of *Defusing Hate: A Strategic Communication Guide to Counteract Dangerous Speech*, a 2014 Genocide Prevention Fellow at the United States Holocaust Memorial Museum's Simon-Skjodt Center for Prevention of Genocide, and the Founder and former CEO of Sisi ni Amani Kenya.

Laura Livingston is the Programs Manager at Over Zero. She previously advised and managed civil society human rights, transitional justice, and rule of law programming in the Balkans, Sri Lanka, and East Africa.

Toda Peace Institute

The **Toda Peace Institute** is an independent, nonpartisan institute committed to advancing a more just and peaceful world through policy-oriented peace research and practice. The Institute commissions evidence-based research, convenes multi-track and multi-disciplinary problem-solving workshops and seminars, and promotes dialogue across ethnic, cultural, religious and political divides. It catalyses practical, policy-oriented conversations between theoretical experts, practitioners, policymakers and civil society leaders in order to discern innovative and creative solutions to the major problems confronting the world in the twenty-first century (see www.toda.org for more information).

Contact Us
Toda Peace Institute
Samon Eleven Bldg. 5th Floor
3-1 Samon-cho, Shinjuku-ku, Tokyo 160-0017, Japan
Email: contact@toda.org

[i] Jonathan Leader Maynard, "Rethinking the Role of Ideology in Mass Atrocities," *Terrorism and Political Violence*, 26(5), 821-841 (2014).

[ii] Mikolaj Winiewski, Karolina Hansen, Michal Bilewicz, Wiktor Soral, Aleskandra Swiderska, Dominika Bulska, "Contempt Speech Hate Speech" (2018), http://www.ngofund.org.pl/wp-content/uploads/2017/02/Contempt_Speech_Hate_Speech_Full_Report.pdf.

[iii] The United Nations' Convention on the Prevention and Punishment of the Crime of Genocide criminalised "direct and public incitement to commit genocide." Convention on the Prevention and Punishment of the Crime of Genocide, Dec. 9, 1948, 102 Stat. 3045, 78 U.N.T.S. 277. The statutes for the International Criminal Tribunal for Rwanda (ICTR) and the International Criminal Tribunal for the Former Yugoslavia (ICTY) similarly criminalise instigating, ordering, or otherwise aiding and abetting in the planning or execution of crimes listed in the statute. Statute of the International Tribunal for Rwanda, art. 6.1., Nov. 8, 1994; Statute of the International Tribunal for the Prosecution of Persons Responsible for Serious Violations of International Humanitarian Law Committed in the Territory of the Former Yugoslavia Since 1991, art. 7.1, May 25, 1993.

[iv] Susan Benesch coined the term dangerous speech. Susan Benesch, "Countering Dangerous Speech: New Ideas for Genocide Prevention," *U.S. Holocaust Memorial Museum*, 5, https://www.ushmm.org/m/pdfs/20140212-benesch-countering-dangerous-speech.pdf; Susan Benesch, "The New Law of Incitement to Genocide: A Critique and a Proposal," https://dangerousspeech.org/new-law-of-incitement-to-genocide/; Jonathan Leader Maynard and Susan Benesch, "Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention," *Genocide Studies and Prevention: An International Journal* 9, no. 3 (2016).

[v] The role of Twitter and Facebook in contributing to the Arab Spring movements has been widely discussed. A former National Security Council adviser even nominated Twitter for the Nobel Peace Prize for its role in the Arab Spring. Mark Pfeifle, "A Nobel Peace Prize for Twitter?" *The Christian Science Monitor*, July 6, 2009, https://www.csmonitor.com/Commentary/Opinion/2009/0706/p09s02-coop.html. Cassette tapes with recorded sermons of Ayatollah Khoeimini, then exiled in a Paris suburb, were smuggled into Iran and contributed to growing support for the Iranian Revolution. A former Iranian government official even acknowledged cassettes to be "stronger than fighter planes." Stephen Zunes, "Iran's History of Civil Insurrections," Huffington Post (July 2009), https://www.huffingtonpost.com/stephen-zunes/irans-history-of-civil-in_b_217998.html**.**

[vi] In Rwanda, radio broadcasts were found to have contributed to the genocide without "a firearm, machete, or any physical weapon." Prosecutor v Nahimana, Case No. ICTR 99-52-T, Judgment and Sentence, para 972, 1099, (Dec. 3, 2003). Also consider how Joseph Goebbels, Hitler's chief propagandist, relied on videos to reinforce pro-Nazi ideologies. Jay W. Baird, "From Berlin to Neubabelsberg: Nazi Film Propaganda and Hitler Youth Quex," Journal of Contemporary History 495 (1983); Leonard W. Doob, "Goebbels' Principles of Propaganda," Public Opinion Quarterly, 428 (1950).

[vii] The Peace Factory started the "Friend me 4 Peace" campaign, facilitating individuals across conflict lines to friend one another to express a shared desire for peace. Through a Facebook page, individuals can opt-in to be "friended 4 peace." Peace Factory then facilitates other group members to friend that individual.

[viii] Rebecca Hersher, "What Happened When Dylann Roof Asked Google for Information About Race?," NPR (Jan. 10, 2017), https://www.npr.org/sections/thetwo-way/2017/01/10/508363607/what-happened-when-dylann-roof-asked-google-for-information-about-race.

[ix] Soroush Vosoughi, Deb Roy, Sinan Aral, "The spread of true and false news online," Science (Mar. 9, 2018), http://science.sciencemag.org/content/359/6380/1146; Max Fisher and Amanda Taub, "In Search of Facebook's Heroes, Finding Only Victims," NYTimes (Apr. 22, 2018), https://www.nytimes.com/2018/04/22/insider/facebook-victims-sri-lanka.html; Max Fisher and Amanda Taub, "How Everyday Social Media Users Become Real-World Extremists," NYTimes (Apr. 25, 2018), https://www.nytimes.com/2018/04/25/world/asia/facebook-extremism.html; Joshua Bielberg and Darrell M. West, "Polit-

ical polarization on Facebook," Brookings (May 2015), https://www.brookings.edu/blog/techtank/2015/05/13/political-polarization-on-facebook/; Max Fisher and Amanda Taub, "Where Countries are Tinderboxes and Facebook is a Match," NYTimes (Apr. 21, 2018), https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html.

[x] Joshua A. Tucker et. al., "Social Media, Political Polarization, and Political Disinformation: A Review of Scientific Literature," Prepared for the Hewlett Foundation (Mar. 2018), https://hewlett.org/wp-content/uploads/2018/03/Social-Media-Political-Polarization-and-Political-Disinformation-Literature-Review.pdf.

[xi] Max Fisher and Amanda Taub, "In Search of Facebook's Heroes, Finding Only Victims," NYTimes (Apr. 22, 2018), https://www.nytimes.com/2018/04/22/insider/facebook-victims-sri-lanka.html.

[xii] Jason Patinkin, "How to Use Facebook and Fake News to Get People to Murder Each Other," *Buzzfeed News*, Jan. 16, 2017, https://www.buzzfeed.com/jasonpatinkin/how-to-get-people-to-murder-each-other-through-fake-news-and?utm_term=.mdl2xvQlQ#.xb1QaZk1k.

[xiii] Amanda Taub and Max Fisher, "Where Countries are Tinderboxes and Facebook is a Match," New York Times (Apr. 21, 2018), https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html.

[xiv] Friesen, J., Campbell, T., & Kay, A. (2014). The Psychological Advantage of Unfalsifiability: The Appeal of Untestable Religious and Political Ideologies. *Journal of Personality and Social Psychology.*

[xv] Monica Pivetti & Marina Camodeca, "Shame, Guilt, and Anger: Their Cognitive Physiological, and Behavioral Correlates." *Current Psychology* (June 2015).

[xvi] These insights are drawn from: Nyhan and Reifler, "Misinformation and Fact-Checking: Research Findings from Social Science," New America Foundation.

[xvii] C. McCauley and S. Moskalenko, "Mechanisms of Political Radicalization: Pathways Toward Terrorism," (2008).

[xviii] Elizabeth Levy Paluck, "Reducing Intergroup Prejudice and Conflict Using the Media: A Field Experiment in Rwanda," Journal of Personality and Social Psychology (2009).

[xix] Sinan Aral, "The Problem with Online Ratings," MIT Sloan Management Review (Dec.2013), https://sloanreview.mit.edu/article/the-problem-with-online-ratings-2/.

[xx] See, for example, the "I Am Here" movement in Europe: https://www.dw.com/en/german-anti-hate-speech-group-counters-facebook-trolls/a-38358671

[xxi] See: Leyens, J. Ph., Paladino, M. P., Rodriguez, R. T., Vaes, J., Demoulin, S., Rodriguez, A. P., & Gaunt, R. (2000). "The emotional side of prejudice: The attribution of secondary emotions to ingroups and outgroups". *Personality and Social Psychology Review*. **4** (2): 186-197; Vaes., J; et al. (2004). "On the behavioural consequences of infra-humanization: The implicit role of uniquely human emotions in intergroup relations". *Journal of Personality and Social Psychology*. **85**: 1016-1034; Cuddy, A., Rock, M., & Norton, M. (2007). "Aid in the aftermath of Hurricane Katrina: Inferences of secondary emotions and intergroup helping". *Group Processes & Intergroup Relations*. **10**: 107-118.

[xxii] Amy J.C. Cuddy, Susan Fiske, et al., "Stereotype Content Model Across Cultures: Towards Universal Similarities and Some Differences," Br. J. Soc. Psychol. (Feb. 2014), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3912751/.

[xxiii] Kteily, N., Hodson, G., & Bruneau, E. (2016). They see us as less than human: Meta-dehumanization predicts intergroup conflict via reciprocal dehumanization. *Journal of Personality and Social Psychology*, 110(3), 343. https://pcnlab.asc.upenn.edu/20161-they-see-us-as-less-than-human/.

[xxiv] This is the case in particular for high-prejudice in-group members (see: Vorauer, J. D., & Sasaki, S. J. (2009). Helpful only in the abstract? ironic effects of empathy in intergroup interaction. Psychological Science, 20, 191-197.)

[xxv] Emile Bruneau, Mina Cikara, Rebecca Saxe, "Minding the Gap: Narrative Descriptions about Mental States Attenuate Parochial Empathy" (Oct. 2015). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0140838.

[xxvi] Elizabeth Levy Paluck, "Reducing Intergroup Prejudice and Conflict Using the Media: A Field Experiment in Rwanda," Journal of Personality and Social Psychology (2009).