Toda Peace Institute

Policy Brief No. 38 April 2019

Mapping Responses to Social Media Threats

Lisa Schirch

Introduction

Recognizing the global relevance of social media threats to human rights, democracy, and peace, the Tokyo-based Toda Peace Institute convened twenty experts from the fields of peacebuilding, democracy, governance, and human rights for an international workshop in December 2018. This workshop is part of a larger programme at the Toda Peace Institute on social media and peacebuilding which includes a series of policy briefs, beginning with a peacebuilding review of the "Social Media Impacts on Social and Political Goods" in October 2018. The Toda Peace Institute is planning further policy briefs and workshops, while a "Global Summit on Technology and Peacebuilding" is planned for 2020.

Alarming stories about social media hit the news almost every day with headlines announcing violations of user privacy as social media companies collect and sell our personal information; Russian troll farms attempting to suppress the African American vote in the US election to assist the Trump campaign; and Facebook executives delaying, denying, and deflecting responsibility for the vast impacts of their platform on democracy, rights to privacy, polarisation and personal safety.

While techno-optimists imagined technology connecting and bringing people together around the planet, new technologies have come with unimaginable costs to our privacy, our democracy and societal relationships, and to peace and human security. Social media threats to society amplify the pace of polarisation and hate speech but did not invent these age-old problems. These problems are too big for the tech industry to fix on its own.

This report identifies social media threats on social cohesion, human rights, violence and democracy and then identifies creative options for addressing those threats through:

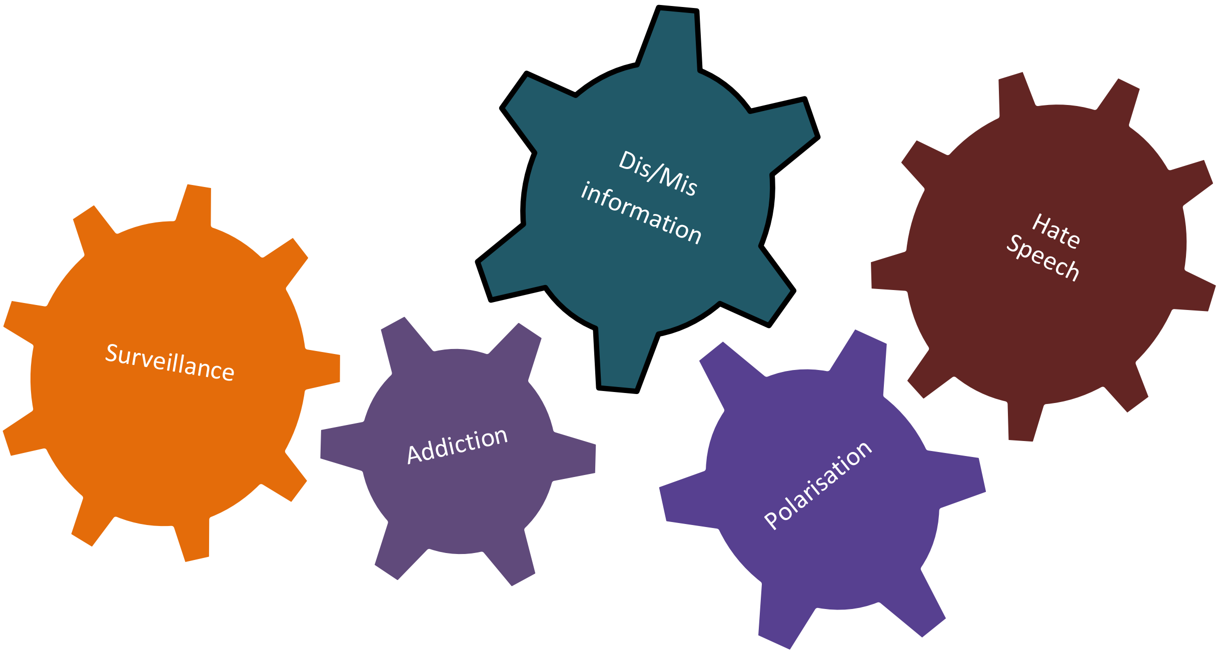
- Building a better bridge between offline dialogue and online platforms.
- Helping tech companies improve their platform design and moderation.
- Supporting civic tech and peace tech options for addressing social media threats.

- Mobilizing civil society to develop campaigns to address social media threats directly or through leveraging pressure on tech companies and governments.
- Leveraging financial and legal pressure on tech companies.
- Recognizing the education and research necessary to develop better long-term solutions.

Social Media Threats

There are at least five interrelated problems related to social media.

Surveillance: Social media platforms make money by providing free communication platforms for people in exchange for extracting information on users. Advertisers use this information about where users live, what they like, what they believe and so on, to target specific ads to specific people. Some call this "surveillance capitalism" where personal information is a valuable resource. Companies like Facebook and Google which provide communication and information services to over a third of humanity make their profit from information gathered through this surveillance. These tech companies are distinct. They do not sell a product to consumers. Their product is consumer data. Their profits depend upon increasing the number of people on their platforms, nudging people to share more information, spend more time online, and get more of their news from the Internet.



Addiction: Social media addictions can affect users of all ages. Social media platforms use psychological research to create positive rewards through colour and social affirmation, such as the Facebook "Like" shiny blue button that seems to offer the same impact on the brain as do some drugs.

Dis/Misinformation: The psychological temptation to spend more time on social media has consequences for the quality of information we consume. On social media, anyone can publish anything. Real news mixes with deceptive propaganda. Some studies suggest false information, exaggerated information, and emotionally alarming information spreads more quickly than other news because it triggers the "lizard" brain and gains more views and clicks. Popularity then translates to legitimacy as algorithms spread this "information pollution" including both disinformation (intentionally fake news) and misinformation (fake news that a user shares without intending to deceive).

Polarisation: Information pollution impacts how people view others. Social media algorithms might be creating "echo chambers" where people receive information that reinforces their worldviews and insulates people from hearing other people's experiences and points of view. Online echo chambers may result in further polarisation, with people lacking understanding of how other's view the issues, and emotionally distancing themselves from others.

Dangerous Speech: For at least some percentage of people, polarisation seems to contribute to hate speech (saying things that dehumanise others) or dangerous speech (making direct threats to others). Social media posts expressing fear, anger, and hate draw more attention, and create more profit for advertisers on social media. Social media may also create safe havens for hate speech and extremist rhetoric, which go unchallenged because they happen out of the public eye.

Like gears in a machine, each of these problems spur on other problems. Addiction impacts how much time people spend online, increasing the likelihood that they will be exposed to disinformation and misinformation increasing the impact of online disinformation, potentially heightening polarisation tied to algorithm-driven filter bubbles, and permitting hate speech to spread more quickly. Acknowledging the interconnectedness of surveillance, addiction, disinformation, polarisation, and hate speech, the Toda Peace Institute gathered experts to find solutions.

Analyzing Threat Sources

The threats in social media come from diverse sources, some with economic goals and some with political goals. On a mass scale, state-run "troll farms" or "web brigades", particularly from Russia and China, but also Western countries, manipulate information online. Robots or "bots" flood the internet with stories and memes intended to deceive, confuse and distract people, aiming to undermine democracy and stability all over the world. As former Ebay executive Colin Rule notes, "We had fraudsters who went to work every day in suits, in skyscrapers, with Ph.D.s in computer science, who were trying to defraud our users."[1] A second category of threat is the violent extremist groups in many countries who recruit new members online. This category includes white nationalist extremists in North America and Europe, as well as extremists elsewhere. Authoritarian elected governments can advocate

[1] https://www.mediate.com/articles/Colin-Rule-Leaving-Facebook.cfm

violent extremist ideas such as advocating direct violence against Muslims in Myanmar, India, Sri Lanka, and the US or offering free reign to corporations executing Indigenous leaders and clearcutting the Amazon forest in Brazil. A third category of threat is that the average citizen spreads misinformation, hate speech or amplifies polarisation on social media. Individuals and small organizations may do this for economic gain or "click bait" or to advance a political agenda. Options for addressing social media threats need to take into consideration these different actors and their motivations.

Assessing Stakeholder Interests

There are four broad sets of stakeholders concerned about social media threats: the public, governments, financial sector, and the tech companies. Each set of stakeholders holds different interests, illustrated in the chart below.

Civil Society/ Public	**Governments**	**Finance Sector**	**Tech Companies**
• Privacy • Communication • Information	• Communicate with citizens • Protect state interests • Regulation or Control	• Profit • Reputation • Stable economic markets	• Profit through user engagement • User demands • Reputation • Resisting regulation • Social goods

Civil society and the broader public largely care about their privacy, their ability to communicate with people to maintain and extend relationships, and sharing or receiving information relevant to their lives. Governments care about social media because it affects how they communicate with citizens, and it relates to their state interests. Some states aim to protect citizens from disinformation or cyber threats, while others seem to use social media to threaten citizen interests. The finance sector cares about social media threats as they pertain to the reputation and profits of tech companies, the insurance companies that protect them, and the impact they have on economic activity and perceptions of stability.

There are at least five motivations driving tech companies as they respond to social media threats. The business model depends on *user engagement, reputation, and user demands.* Advertisers pay platforms based on user engagement in a platform and the quality of data that the platform extracts from its users and provides to advertisers to target audiences for their products and messages. Investors want platforms that increase user engagement, as measured by how much people share and respond, how much time is spent on the platform, how many users have accounts, and how often they log on to the platform. Social media companies also have a goal to *limit government regulation* of their industry, as they believe that regulation would threaten profitability. At the same time, regulation tends to favour industry incumbents, so regulation may, ironically, make it more difficult for startups, including competitive social media platforms, to survive. Social media companies care about their reputation and want to limit stories that harm public perceptions of their company. Related to this goal, social media companies want to contribute, in general, to the *social good*, especially their stated goal of connecting people. But social media companies are a far cry from having a triple bottom line that values people, planet, and profit. Surely shareholders would not just sit by as social media profitmaking threatens people around the planet. Options for addressing social media threats need to take these factors into consideration.

Solutions

There are no quick fixes to social media threats. There are wide disagreements and conceptual contradictions on how to define disinformation, hate speech, privacy, and addiction. Efforts to stem disinformation or hate speech can also censor legitimate concerns about human rights.

The group identified a range of potential options for different stakeholders to address interrelated threats. These include the following, detailed in more length below:

- Building a better bridge between offline peacebuilding and social media technology.
- Helping tech companies improve their platform design and moderation.
- Supporting civic tech and peace tech options for addressing social media threats.
- Mobilizing civil society to develop campaigns to address social media threats directly or through leveraging pressure on tech companies and governments.
- Leveraging financial and legal pressure on tech companies.
- Recognizing the research necessary to develop better long-term solutions.

Peacebuilding and Social Media Technology

Social media did not create the problem of polarisation or conflict. Peacebuilders have been addressing these issues for decades. Now, the field needs to step up not only to leverage social media capabilities for improving problem solving and relationships, but to address wide ranging social media threats that amplify ancient problems facing humanity.

Peacebuilders know that listening is transformative. People are more likely to be open to hearing other points of view (aka “attitude complexity”), feel less anxious, and be more self-aware when feeling acknowledged and understood by a good listener. As one example, see the Harvard researchers exploring the power of listening in helping people to change.[2]

Dialogue processes offer a kind of “spa” for being heard and acknowledged. In terms of polarisation, changing the norms and tone of conversations on social media is necessary. The goal is not to change people’s minds on the issues (issue polarisation) but to change how they talk and feel about people who see issues differently (affect polarisation). Some peacebuilding groups like Search for Common Ground[3] have already held conversations with Facebook and other social media platforms to increase their understanding of dialogue efforts such as Soliya[4] and the “common ground approach.” Similarly, the Human Library[5] offers an opportunity to dialogue with diverse real people who are on “loan.” The programme

[2] https://hbr.org/2018/05/the-power-of-listening-in-helping-people-change
[3] https://www.sfcg.org/what-we-do/
[4] https://www.soliya.net/
[5] http://humanlibrary.org/

encourages readers not to judge a "book" by its cover and to challenge stereotypes and prejudices.

In a global community divided by politics, economics, language, and conflicts, technology is a significant factor for peace and security. In the short term, social media can amplify the power of dialogue and intergroup understanding if we can appreciate the complex new reality of communication, and then build better bridges between online and offline relationships and conflict transformation processes. People need to see dialogue modeled and witness the transformation that people describe from being heard and acknowledged. Social media can strengthen dialogue through video clips capturing people's experience of conversation and invitations for people to join in a social media discussion. These tactics can discourage people from demonizing others and offer inspirational ideas for how people can listen to and respect one another.

In the long term, the links between the fields of peacebuilding and technology need more robust exploration.

1. **A Global Summit on Peacebuilding and Technology** illustrates one example of a way to build out this agenda. The Toda Peace Institute could convene an off-the-record global summit with key technology, finance, peacebuilding and government partners and leaders to identify key research and lessons learned while establishing the relationships and leverage needed to foster real changes. High-level stakeholders such as Microsoft, Hewlett Packard (which helped to establish the conflict resolution field), Japanese firms and equipment providers, individuals like Bill Gates, Pierre Omidyar, and Desmond Tutu, and civil society groups such as Civic Signal and ICT4Peace, for example, would gather with peacebuilding leaders and those actively working on product and feature design—who would be more essential than policy heads in implementing changes. A summit would require high-level support, as well as ongoing working groups and staff to continue the conversation and develop a strategy for monitoring progress and success. The project could spotlight tech people focused on the social good and amplify their voices.

2. **Tech Sector:** Within the tech sector, materials, reports, and training could help to build capacity to think about social cohesion, polarisation, civic engagement, and broader peacebuilding themes. Training in conflict assessment, conflict sensitivity, social cohesion, and "do no harm" could help the tech industry prevent unintended impacts by more thoroughly understanding local context. At a more basic level, computer science programmes could incorporate more courses in ethics related to the social and political impacts of technology.

3. **Peacebuilding and Human Rights Sector:** Within the field of peacebuilding and human rights, there is a need for greater awareness and understanding of technology, as well as the impact of social media on their work. The Alliance for Peacebuilding could host a working group to bring together its network on these issues. Early response

teams in countries with violent conflict could prepare to de-escalate social media-related violence. Organizations could train individual fact-finders using existing modules from Berkeley's Human Rights Investigation Lab[6].

4. **Key Opinion Leaders:** Government, education, religious, and civil society leaders may benefit from greater social media literacy, specifically an appreciation for how social media relates to issues such as polarisation, hate speech, and violence.

5. **Kids and the General Public**: Social media literacy is necessary to foster a collective critical consciousness of social media content. Such literacy can take the form of national programmes, radio spots, television spots, and public service campaigns on topics such as responding to fact checking, regulating emotions, confronting hate speech, and depolarising by listening and building rapport before seeking to persuade. Activists can teach people how to identify and combat disinformation, like the NGO community's InterAction Disinformation Toolkit.

Platform Design

The design of social media platforms reflects the aims of company shareholders: to get people to engage and share on social media longer so the company can collect more information on each user and increase advertising revenue. A plethora of options exist to redesign social media platforms to better serve humanity. While a few groups are testing these ideas, more research is necessary to determine the effectiveness – and the unintended impacts – of such changes.

Efforts to redesign social media platforms address an array of concerns that stem from the current design of the platform and algorithms, which seem to amplify and reward addiction, disinformation, polarisation, and hate speech. Platforms should protect people's minds from addiction, social comparison, anxiety, and trauma spread through online hate. Platforms should incentivise positive social relationships, truth, deliberation, and respect. Platforms should have guidelines for use, but more importantly establish norms that encourage users to model positive civic engagement.

The goal would be to slow down discussions, recognising that democracy and deliberation require time. Researchers find that highly emotional material spreads more quickly on social media than material that is not highly emotional. Disinformation campaigns take advantage of that fact by spreading fake news or news framed in highly emotional ways that people will share. From a neuroscience point of view, social media could expose people to information that enables better deliberation and helps people to embrace the ambiguity and complexity of events. Platforms could encourage affirmation of people who are respectful to others, and it could reward people who have a positive reputation for the way that they treat others on social media. Since most people do not have exposure to good conflict resolution skills, platforms need to help people model these behaviors for others.

[6] https://www.law.berkeley.edu/research/human-rights-center/programs/technology/human-rights-investigations-lab-internships/

Below are suggestions by workshop members about platform design:

1. **Change algorithms:** Current algorithms aim to provide people with content that they want to see. Democracy and social cohesion require encountering information that surprises us, gives us new information, or challenges us to think in new ways. Instead of algorithms creating "echo chambers", platforms can encourage interacting with different ideas and people.

2. **Change comment sections:** While social media platforms excel at creating ways for people to share their opinions, they are not built to help people listen to and acknowledge one another. Listening is a powerful tool for depolarising groups. Comment sections on social media could promote community by offering better ways for people to take turns being heard and understood, ensuring that temperate first posts prevail in news posts, and offering predefined emoticon buttons.

3. **Gamify platforms:** Many people like online tests and games. Competitions could be set up to see who has the most diverse social network, who posts the most diverse types of information, who cultivates online norms successfully, who promotes accuracy, etc. Social media companies could offer a reward, such as free services or products, to those who generate positive online culture.

4. **Improve verification on platforms:** Social media companies could improve strategies to identify disinformation, dispel myths, and protect people from information pollution. Verification also could be used to address fake accounts by improving strategies for verifying users, especially high-profile ones (see the blue tick on Twitter as an example). Companies could create "read" receipts to minimise anonymity.

5. **Colour-Code Information Sources:** Social media platforms mix together a wide variety of information about individuals, friends, family, memes, fake news, and genuine news. One option for addressing this confusing mix of information is to offer contextual clues by marking information sources with colour-coded boxes that indicate reputable news, opinions, family and friends, bots, and paid advertisements.

6. **Create flags:** Social media platforms could incorporate a feature that detects inflammatory speech and non-word hate speech in the form of emojis or memes. Platforms also could create a metric to grade levels of inflammation and danger of posts or identify ways to address gender differences in contributing to dangerous speech.

7. **Offer pop-up options:** Flagged material could trigger a pop-up box offering a prompt for the person to pause and reflect on their choice. A pop-up could ask "Do you want to post this?", provide alternative phrases to hateful choices, encourage productive speech, and prompt reflection through follow-up questions. A pop-up could offer a link to training on disinformation or communication skills. A pop-up also could provide links that offer dialogue or volunteer opportunities for positive involvement on a divisive topic when a user posts related content.

8. **Add buttons or tags related to social cohesion:** Social media platforms incentivise engagement through buttons and tags that allow people to interact with material. Facebook expanded its single "like" button with other options, including "love", "sad",

"wow," and "angry." New ideas range from a symbol like a "bridge" that could spotlight depolarising actors; a flame to signal resilience, hope, or transformation; and a symbol to affirm respectful content.

Platform Moderation

Changing the nature of online communication requires a combination of bottom-up and top-down interventions involving moderation and social influence.

There are two types of problems that require moderation on social media. The first is individual actors who use hate speech, make threats, and create division online by spreading disinformation. The second is paid trolls and bots that spread large amounts of disinformation aimed at undermining facts, causing the public to distrust reputable news sources, distracting from real news, and sowing confusion and chaos.

Below are suggestions by workshop members about platform moderation:

1. **Invest in Online Norms:** The best way to change the quality of conversations on social media is to create common norms, a culture in which groups all pull each other back into line. Rules can be important too, but more as a last resort. Extremism usually starts with name-calling and dehumanizing other groups. It is difficult to regulate that type of speech. Social media companies can invest more in developing shared norms and creating ground rules for user behavior. Facebook has training and moderator communities, but the emphasis is on keeping rules, not on helping to foster civic norms. Sometimes it is just a small group of three to five people who can change the tone of online conversations. Rules should be a last resort, prioritised only if norms have broken down.

2. **Provide community moderation:** When people have shared norms, such as avoiding name-calling, people that stray off the path can be pulled back into a pattern of civil discourse by others. A more robust approach could be volunteer teams of 10 people or online moderators who are trained in evidence-based methods to diffuse a potentially dangerous post. Certification or a moderator tag might denote someone's status as an online volunteer moderator.

3. **Pay staff moderators who speak local languages:** Current moderation is based on the number of reports made in a context. In most countries, social media companies do not have many staff devoted to responding to complaints. It can take days, weeks, or even longer for social media platforms to respond to complaints and requests to remove information. Tech companies need to expand short-staffed "harm teams" and increase staff moderation. Often, there are few moderators who speak the local languages, understand local terms used as hate speech, or know the political context enough to identify a potential threat. Tech companies could partner with local actors and on-the-ground networks to create accurate "slur lists" and a local lexicon of hate speech, monitor hate speech, verify posts, and spread knowledge to correct false information.

4. **Moderator bots, algorithms, and tech solutions:** There exist various tech solutions to the problems of online hate speech, disinformation, and polarisation. More efforts are

needed to bring together local human rights activists, peacebuilding and civic engagement experts, and tech companies to assess how to improve moderation tech options.

5. **Provide transparency in tech moderation rules:** Social media companies must strike a balance between preventing hateful posts and allowing personal expression. Touting free speech, social media companies have been hesitant to invest staff in moderation. And human rights activists censored on social media by moderators press companies to allow more freedom of speech, especially for posts that critique governments. Tech company choices on how and how much to moderate online information impacts democratic processes and can be a matter of life or death if a post calls for violence against an individual or group.

6. **Create crisis-response hotlines:** Tech companies need to improve "escalation mechanisms" for local actors to get the attention of tech decision makers through a hotline for dangerous speech. A hotline, for example, could request a platform slowdown in extreme situations, or provide immediate moderation to a life-threatening situation.

7. **Increase actor accountability and disincentives for bad actors rather than focusing on content:** Moderators often deal with specific posts or content rather than the accounts of those posting it. A relatively small group of people are often responsible for online bullying, hate speech and threats. Finding ways of sanctioning these "bad actors" on social media could have an impact on disinformation, polarisation, and hate speech.

8. **Combat context collapse**: "Context collapse" happens when people post photos, stories, or information without indicating the context in which this content emerged. In Myanmar, for example, the military spread photos of dead people, falsely accusing Rohingya Muslims of killing them. In the US, people spread clips of speeches by people without the full context of the communication, such as when Hilary Clinton gave a speech about closing West Virginia coal mines and investing in new jobs but social media pages included just the clips about closing mines. Another issue is the threat of "deep fakes" in which social media use high tech to falsely portray a person saying or doing something.

9. **Research and create guidance on moderation best practices.** The online dialogue group Smart Politics[7] teaches progressive US activists how to change hearts and minds through their Radical Conversations[8] methodology. Smart Politics Founder Karin Tamerius describes how the organization taps people who seem to have moderation skills and invites them to be "Ninja" moderators who practice conflict resolution skills on unwitting commenters. Peacebuilding programmes such as Soliya[9] and The Commons[10] also experiment with depolarisation efforts online.

[7] https://www.joinsmart.org/

[8] https://www.joinsmart.org/the-rcc.html

[9] https://www.soliya.net/

[10] http://howtobuildup.org/wp-content/uploads/2016/04/The-Commons-A-pilot-methodology-for-addressing-polarization-online-2-27-18.pdf

Civic/Peace Tech

Peace tech and civic tech are broad categories of technology designed to improve social and political goods, including improving relationships between state institutions and communities, checking facts, spreading truthful news, and monitoring human rights and peace accords.

Civic/peace tech offers a variety of apps and platforms to address issues. There are platforms and apps that focus on civic data and transparency. There are platforms for depolarisation like My Country Talks[11] and Buzzfeed Outside your Bubble[12]. The "Civic Tech Field Guide"[13] identifies 200 subcategories of civic tech. PeaceTech Labs[14] and PeaceHack[15] bring together tech experts with peacebuilding practitioners to create new tech-based mechanisms for dialogue, maps of conflict indicators, and crisis response networks to defuse violence.

Civic/peace tech can also address social media threats. A variety of tech tools already assist in fact checking, accountability, and human rights reporting. The Coral Project[16], founded as a collaboration between Mozilla, *The New York Times*, and *The Washington Post,* offers free software and guidance to bring journalists and the communities they serve closer together. Meedan[17] verifies breaking news online and helps groups translate and share information online. Meedan helped to create Popup News Room[18], which supports journalists and civil society as they collaborate on news sourcing, and the Credibility Coalition[19], which promotes better standards for online content. Tech for addressing disinformation includes Fact Check[20] and Factmata.com[21], which counter disinformation and spot malicious bots.

Civic/peace tech has an important role in testing out ideas to improve platforms and online engagement that can benefit social and political goods, and to combine online and offline initiatives in which tech interacts with the real world. The Syrian Archive is a Syrian-led collective of human rights activists dedicated to curating visual documentation of human rights violations and other crimes committed by all sides during the conflict in Syria, with the goal of creating an evidence-based tool for reporting, advocacy, and accountability purposes. Offline training helps to expand human rights tech skills. Berkeley's Human Rights Investigations Lab[22] is training the next generation of students in how to find, verify, and analyse social media information—whether photos, videos, or posts—about some of the most pressing human rights challenges of our times. With Reuters, the Berkeley Human

[11] http://www.mycountrytalks.org/
[12] https://www.buzzfeed.com/bensmith/helping-you-see-outside-your-bubble?utm_term=.heOZY8PXel
[13] https://civictech.guide/
[14] https://www.peacetechlab.org/
[15] https://peacehack.io/
[16] https://coralproject.net/
[17] https://meedan.com/en/
[18] https://popup.news/
[19] https://credibilitycoalition.org/
[20] https://www.factcheck.org/
[21] https://civictech.guide/listing/factmata-com/
[22] Above, n 6.

Rights Center published *Hatebook: Why Facebook is losing the war on hate speech in Myanmar.*[23]

While these options offer interesting and constructive tech uses, none of them is employed at a scale large enough to combat the negative uses of existing platforms such as Facebook used by over two billion people.

Civic/peace tech can experiment and research various ways of addressing social media threats, such as the following research ideas:

- Create an online "Social Cohesion Index" that would rate social media platforms and countries.
- Develop and experiment with setting standards for conduct, defusing hate speech, and empowering users to engage with disinformation and related themes.
- Build taxonomies for polarising content.
- Experiment with hate speech interventions.
- Build an online "Listening Corps" that can consist of bots or people modeling active listening.

Advocacy Campaigns

Many human rights, democracy, and peace groups around the world recognise the need to mobilise people power to leverage pressure on social media companies to stand accountable to public interests.

A variety to campaigns are already underway. Microsoft partnered with a handful of civil society groups like Civicus to organise the Digital Peace Now[24] online campaign expressing opposition to technology as a weapon to carry out cyberwarfare with a vague demand, not calling for any specific action but stating, "Our online community must not be a battlefield. We demand digital peace." The #ReformFacebook[25] campaign called for "core institutional reforms to Facebook's Board of Directors to be more independent, accountable of senior leadership, and capable of understanding the civil rights and privacy implications for how its platforms are used."

Avaaz, a global online platform for digital activism, launched a "Fix Facebook"[26] campaign in 2019 with a personal call to Facebook CEO Mark Zuckerberg. It reads, "As global citizens, we call on you to immediately set up effective global systems to delete fake accounts, identify and take down hateful or false content fast, and prioritise content from trustworthy sources. Facebook is undermining our democracies. It's up to you to stop this danger."

[23] https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/?utm_source=twitter&utm_medium=Social

[24] https://digitalpeace.microsoft.com/

[25] https://www.muslimadvocates.org/reform-facebook/

[26] https://secure.avaaz.org/campaign/en/fix_facebook_40/

The “Security Pledge”[27] Campaign is an existing campaign run by US-based civil liberties and human rights groups to address the threats from data collected by tech companies that can be used to harm social groups and interfere with democracy and human rights. The campaign asks tech companies to pledge to secure privacy by building a “surveillance-resistant web” that can stop authoritarian tools and abuses of data.

Effective campaigns involve coordination between a variety of groups to form a movement that includes strategists, communicators, and connectors who identify relevant and clear “asks,” provide analysis and information on the problem, and grow the network of those involved to leverage “people power.” Ideally, campaigns appeal to the broadest possible coalition of groups. Our group developed a variety of concepts for potential future campaigns. These include the following:

1. “**Digital Neighborhood Watch**”: This campaign would aim to change the tone of comments at the community level on particular platforms such as the local newspaper, TV station, and Facebook pages. Many communities have a small group of local “trolls” who have a disproportionate impact on community life. Others have ceded online territory to these trolls because dynamics became so ugly online. In this campaign, a community group, perhaps 25 people from various walks of life, would gather in an open space format every three or four months to design their own digital neighborhood watch programme based on local culture and needs. Members of this neighborhood group would receive training if needed and offer each other support as they engage online on certain platforms to set the tone for mature deliberation and attempt to de-escalate hate speech, dis/misinformation, or other polarising content. This initiative could happen in “test” communities around the country that would sharpen the methodology. Test digital neighborhood watch campaigns could share best practices and cultural adaptations as the campaign expands to more cities.

2. “**Digital Resistance Elves:**” The Baltic States have addressed Russian disinformation for much longer than other countries. Volunteer armies of “digital resistance elves” combat disinformation and improve social media literacy. More research is needed to understand how these online campaigns have worked. Ideally, civil society groups in other countries could launch digital resistance campaigns to build a movement of “elves” to combat disinformation “trolls.”

3. **“Social Media Fast”:** Some participants in the workshop brainstormed the idea of creating a global campaign to coordinate a boycott or strike from Facebook, Twitter or other social media platforms. The campaign could identify clear demands in terms of policy, practice, and immediate changes. For example, a “Facebook Fast” might emphasise the monetary worth of an individual user to different platforms and heighten public awareness of three issues: 1) individual addiction and anxiety related to social media; 2) the surveillance capitalism that threatens privacy; and 3) the social and political impacts of social media on society.

[27] https://www.securitypledge.com/

There are various ways that a "fast" campaign could work. First, a boycott could include a simple pledge by users not to click on any ads on the social media platform, as each click increases ad revenue. Users can leverage pressure on advertisers to prod social media companies to change. As part of the campaign, users could take the social media apps off their phone or put their phone screens on black and white instead of colour, which may decrease the appeal. Third, users could pledge to fast from specific social media platforms for a week or month until the company agrees to take specific action. Fourth, the campaign could include the call for users to close their social media account and possibly join a new platform. As Colin Rule posted on his Facebook page, "FB if you do A, B, and, C I will stay online, [sic] if not I will close my account."

Financial and Legal Pressure

Workshop participants reported on a variety of options for using financial and legal pressure to deal with social media threats.

1. **Lawsuits** against social media companies could be based on their contribution to and amplification of threats to society. Just as polluting corporations have to pay taxes toward funds that go to clean up air and water pollution, social media companies could be taxed for their contribution to information pollution, since a functioning democracy requires information.

2. **Taxation** of social media companies could be based on their impact on social cohesion and democracy. Legislators could justify taxation on the grounds that social media platforms spread social and political pollution by undermining social relationships and political institutions. Taxes could be channeled to fund not-for-profit offline news sources, including public-access news, information "trusts," and civic media.

3. **Insurance companies** could be required to calculate and price insurance premiums for social media companies based on the financial impact and risk of chaos, violence, and political instability stemming from social media.

4. **Regulation** of social media platforms could address key threats to privacy, mental health, information, democracy and safety. Yale Law School professor Jack Balkin argues that social media companies need to be treated as "information fiduciaries" and as such need regulation to protect and care for the public's access to accurate information. The business model of most social media platforms creates incentives for companies to limit regulation of their industry. Shareholders and tech companies already have a powerful lobby to influence elected officials around the world. But there is a growing calculus of financial, political, and social impacts of social media on the world, and there is also a strong lobby that supports regulation, taxation, lawsuits, and pressure on insurance companies.

The European Union, and a handful of other governments, are in ongoing deliberation on the regulation of social media. In May 2018, the European Union's General Data Protection Regulation (GDPR) went into effect with the goal of protecting individual privacy and giving control to individuals over their personal data. In September 2018, the European Union released new rules mandating that social media platforms remove terrorist content within one hour of its posting.

French President Emmanuel Macron, Microsoft and other tech companies launched the Paris Call for Trust and Security in Cyberspace. US Senator Mark Warner's *White Paper* on options for government regulation of social media, the United Kingdom's standards on disinformation, and a Digital Social Contract[28] are other examples. Others note that the UN Principles on Business and Human Rights could be applied to the social media industry. Microsoft is calling for a Digital Geneva Convention that would create new international rules to protect the public from state threats in cyberspace.

In particular, regulation could address the following:

a. **Requiring a License to Operate:** Social media companies could be required to acquire an FCC-like "license to operate."

b. **Setting Moderation Standards:** Regulation could include user protections and requirements for hotlines to respond to crises such as inflammatory posts. Regulations could require content moderation that matches the threat level of a given context. Germany, for example, has passed legislation requiring social media companies such as Facebook and Twitter to apply fines for delays in removing content such as dangerous speech. In response to these new laws, Facebook and Twitter have invested a significantly larger number of resources and staff to work on content moderation. Other countries can learn from German attempts to sanction false information.

c. **Requiring a Risk Audit:** Regulation could require that tech companies cooperate on funding a "Risk Audit" for social media technology in every country. Part of the current problem is that the tech community designing social media largely consists of young white men with an education in technology but little understanding of sociology, political science, or the historical-cultural dynamics of particular regions of the world. This ignorance often seems paired with an arrogance of not knowing what they don't know.

Research

Finally, there is a need for a lot more research to analyse the threats, test assumptions, and develop evidence-based options for addressing the crisis. These research ideas, collated by Lydia Laurenson who participated in the workshop, include a variety of themes.

1. How much of the increase in polarisation can reasonably be attributed to social media and/or mobile tech?

[28] https://digitalsocialcontract.net/

   a. What are the specific impacts of increasing homophily, i.e., people sorting into groups based on their similarities?
   b. What do we know about "asymmetric polarisation" and how it works, its root causes, and interventions?
   c. Are there positive impacts of polarisation?

2. What interventions have been tested to decrease polarisation broadly, or to build peace specifically, both online and offline?
   a. Commission a meta-study of research findings on the impact of social media offline, and of the different types of impacts of social media.
   b. Map peacebuilding strategies around polarisation.
   c. How could Microsoft's Digital Civility Index[29] be used more broadly?
   d. Create scorecards that assess depolarising qualities in leaders, politicians or processes.

3. Do different types and genres of social media (or just different platforms) have significantly different impacts and/or require different interventions? What strategies are available for testing different designs on major social platforms?

4. We need more information about cross-cultural digital media given how many powerful social media platforms exist outside the US, and also given how different social media usage can be in different cultures, even when the platforms are the same.

5. What more do we need to know about surveillance economics and how it works, as well as what tech companies are not operating on this model?

6. What do we need to know about the global underground disinformation and propaganda trade? There are vendors as well as individuals selling their services in this world, and there are state actors making alliances and training each other as well. Has anyone mapped it? Who are the major players?

[29] https://www.microsoft.com/en-us/digital-skills/digital-civility?activetab=dci_reports:primaryr6

The Author

Lisa Schirch, PhD is North American Research Director for the Toda Peace Institute and Senior Policy Advisor with the Alliance for Peacebuilding. Schirch is the editor of "The Ecology of Violent Extremism" (2018) co-author of "Synergizing Nonviolent Action and Peacebuilding" with US Institute of Peace (2018). In 2015, Schirch finished a 3 year project coordinating a global network to write a Handbook on Human Security: A Civil-Military-Police Curriculum and set of 40 peacebuilding case studies on Local Ownership in Security. Schirch works with a global network of civil society, government, and U.N. partners devoted to expanding social and political goods.

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