ANALYSIS & NEW INSIGHTS

Disinformation and Identity-Based Violence

Samantha Bradshaw, DPhil | November 2024

Introduction

Digital technologies are accelerating and exacerbating violence and instability globally. Over the past decade, there have been high-profile examples of how disinformation spread via digital technologies has been used to incite violence against individuals and marginalized communities, stifle social movements, and silence press freedom. Many of these coordinated disinformation campaigns do not just rely on falsifiable claims that undermine truth. They also draw on identity-based disinformation—or disinformation that weaponizes racism, sexism, and xenophobia—to operationalize online and offline violence and the linkages between them. Considering the specific characteristics, impacts, and challenges of identity-based disinformation, how might the application of an identity frame help to adapt or advance existing disinformation interventions to mitigate harm?

Rhetoric, Identity, and Targeted Hate Speech: The Case of Myanmar

Beginning in 2012, Myanmar military personnel posed as celebrities, pop stars, and national heroes as they flooded Facebook with anti-Muslim rhetoric and calls for the ethnic cleansing of the Rohingya, a Muslim minority in Myanmar.[1] Dozens of accounts—created as part of a covert military disinformation operation—positioned themselves as independent sources of news and information, disseminating false and hateful content about the Rohingya.[2] Behind the veil of anonymity, government commentators used fake Facebook accounts to systematically spread disinformation, calling the Rohingya and other Muslims "dogs, maggots and rapists, suggest they be fed to pigs, and urge they be shot or exterminated."[3]

The use of social media in fueling the abuse and displacement of more than 700,000 Rohingya compels analysis. At the time, more than 80 percent of Myanmar's population connected to the internet through mobile apps such as Facebook.[4] However, despite the rapid adoption and use of social media in Myanmar, Facebook had only one Burmese-speaking content moderator—who was based in Ireland—to monitor the spread of government disinformation campaigns and hate speech among Myanmar's 1.2 million active users.[5] It wasn't until late 2018, after significant public backlash, that Facebook first removed accounts and pages operated by Myanmar military personnel and onboarded an additional 99 Myanmar-language content moderators.[6] Despite the high-profile attention of this case, as well as the responses adopted by Facebook (now Meta), anti-Rohingya content and hate speech continued to spread on the platform. For example, in March 2022, Global Witness found Facebook's AI-approved advertisements contained hate speech targeting Rohingya.[7]

The Weaponization of Identity

The case of Myanmar helps demonstrate that social media is not an inherently liberating technology but can be weaponized by governments to control the information space, suppress human rights, and incite violence.[8] While high profile examples of coordinated disinformation campaigns often focus on how false narratives and fake accounts might disrupt elections,[9] the case of Myanmar illustrates the unique ways that disinformation can be weaponized to foment fear, hatred, and violence against marginalized populations. Here, disinformation does not just rely on falsifiable claims but also leverages **identity** and the vocabulary of racism, sexism, and xenophobia to operationalize online and offline violence and the linkages between them.

This paper explores the concept of identity-based disinformation and how it is operationalized and weaponized to incite violence, promote affective polarization, and undermine human rights around the globe. The paper invites investigative questions, such as: How does the inclusion of identity into disinformation narratives and frameworks affect the quality of our digital information ecosystem and how individuals and groups are able to access and exercise fundamental human rights? Are different approaches needed to combat disinformation that draws on identity frames? And what can civil society and technologies do better to foster a digital culture of peace that prioritizes inclusivity and respect for human dignity in online spaces?

The Identity and Disinformation Nexus

Disinformation is often defined as the deliberate creation and dissemination of false or misleading information with the intent to deceive or manipulate.[10] Identity-based disinformation is a specific form of disinformation that targets individual or group identities, such as race, religion, ethnicity, gender, or sexual orientation. Like other kinds of "coordinated, inauthentic behaviors,"[11] the purposeful spread and false amplification of identity-based disinformation is not just an individual expression of individual bias but instead represents the systematic

weaponization of discrimination via technology to make hateful narratives go viral.

Identity-based disinformation can be considered distinct from other forms of online harm, like harassment, that also leverage the vocabulary of discrimination and hate. Where the harassment-disinformation nexus involves systematic and coordinated attacks targeting a specific individual,[12] identity-based disinformation shifts the focus to the individual and group levels, where disinformation is designed to be more pervasive across a broader community of people but with individual impacts. Thus, identity-based disinformation creates a model of attack that weaponizes group-level identities in ways that impact not just individuals but entire communities and demographics.

Where disinformation traditionally emphasizes the intentional distortion of facts, the inclusion of identity into the framework of disinformation helps define the ways preexisting prejudices, stereotypes, and biases can also be weaponized for political purposes.[13] At the core of identity-based disinformation is the exploitation of individuals' or groups' senses of identity, belonging, and social standing. An individual's creation of their identity is something that is contextually derived from the unique histories, cultures, and power structures present across different societies and online platforms. As a form of "dangerous speech,"[14] identity-based disinformation can often glorify violence as something that is virtuous and frame it as necessary against perceived threats from a marginalized group.[15] By creating, reinforcing, and exploiting societal divisions through such in-group and out-group narratives, malicious actors can exaggerate perceived differences and conflicts between groups for political gain and amplify the potential for conflict and unrest.

Research on identity-based disinformation has highlighted how such narratives are used in contemporary influence operations to exacerbate perceived differences and conflicts both within and between groups.[16] Crucially, these societal divisions need not be real but can be imagined and then projected onto a group of people through coordinated disinformation tactics. Although identity is shaped by an individual's own unique history, culture, and broader community, perpetrators of violence often conceptualize a reductive form of identity and impose it on their victims. These "essentializing identities"[17] often exploit general stereotypes of nondominant groups in society and might not reflect how individuals actually see themselves nor capture the range of identities individuals might hold. However, by manipulating one identity, perpetrators of violence can justify their actions and further dehumanize a group.

Broad Implications of Identity-Based Disinformation for Security and Democracy

The range of harms resulting from identity-based disinformation are broad and can have immediate impacts on individuals and groups, as well as long term societal impacts. By exploiting existing societal divisions, disinformation campaigns that weaponize

identity can exacerbate tensions, leading to increased hostility and violence against individuals and marginalized groups. In extreme cases, this identity-based disinformation can promote violence, mass displacement, and genocide against communities of people. Additionally, democratic institutions are negatively impacted by identity-based disinformation, which can undermine political mobilization, discourage political participation, threaten press freedom, and increase affective polarization.[18]

First, identity-based disinformation can be used to **stifle political mobilization**. The deliberate use of racially charged, sexist, xenophobic, and homophobic rhetoric not only harbors a culture of intolerance but also serves to harass individuals and activists, stifling social movements and fracturing social cohesion. For example, research on the gender dimensions of foreign-influence operations shows how gender-based disinformation was not only used to harass prominent feminist activists participating in the 2017–2018 US women's marches but also undermine their sense of collective identity and limit their ability to advocate and coordinate.[19] This form of violence can diminish the impact of social movements, ultimately eroding democratic engagement.

Second, identity-based disinformation has **detrimental effects on political participation**, particularly for female politicians. By perpetuating harmful stereotypes and spreading false narratives about women in leadership, these disinformation campaigns create a hostile environment that discourages women's involvement in politics. For instance, during the 2024 elections in Mexico, female candidates faced targeted disinformation that questioned their competence and character, often framing them as unfit for leadership.[20] This not only undermined their campaigns but also contributed to a broader culture that deters other women from entering the political arena. When female politicians are systematically targeted, it stifles diverse representation and reinforces gender inequality in political participation.

Third, identity-based disinformation **can damage press freedom and close civic spaces**. Coordinated campaigns targeting journalists can create a chilling effect, closing civic spaces and deterring investigative reporting. Indeed, many governments have used identity-based disinformation to target journalists through coordinated attacks.[21] While these kinds of attacks can have consequences for all journalists, there are important differences in experience based on gender and race. For example, research conducted by the International Center for Journalists and the United Nations Educational Scientific and Cultural Organization (UNESCO) has found that nearly three out of four female journalists have reported experiencing violence online.[22] Case study analysis of these attacks highlight the important role of identity and gender, where attacks leverage sexism and misogyny and other forms of discrimination such as racism, bigotry, and homophobia to silence female journalists and undermine their credibility.[23] These kinds of attacks can have serious long-term effects on journalists, with impacts ranging from psychological trauma to self-censorship or to women leaving the journalism field.[24] Additionally, when journalists are targeted with identity-based disinformation that characterizes their work as incredible or incites violence against them, broader public discourse can be negatively affected and press freedom chilled.

Fourth, identity-based disinformation significantly **contributes to affective polarization**. Affective polarization is the situation in which individuals develop strong emotional reactions—positive or negative—toward members of their own group compared to those in opposing groups. By disseminating false narratives that exploit existing divisions, these campaigns deepen affective polarization and create an "us versus them" mentality among communities. For example, in the United States, during the 2020 election cycle, disinformation targeting racial and ethnic groups intensified divisions, making it more difficult for individuals to engage in constructive dialogue across political lines.[25] When identity-based disinformation fuels polarization, it not only erodes trust in institutions but also hampers the ability of citizens to find common ground, ultimately undermining the fabric of democratic discourse.

Finally, identity-based disinformation **can exacerbate risks for atrocity crimes**. Ethnic violence and genocide are inherently identity-based crimes.[26] Many governments have used the open nature of digital technologies to facilitate the spread of identity-based disinformation to harm or promote violence against certain populations.[27] For example, in addition to the case of Myanmar, in China, the government has systematically spread identity-based disinformation on social media to vilify ethnic Uyghurs and justify the mass detention of over a million people in so-called reeducation camps.[28] In Ethiopia, Tigrayans have been targeted with identity-based disinformation that has stoked deadly violence and enabled the government's military campaign against the Tigray region.[29] And in Afghanistan, the Taliban has used identity-based disinformation to justify and enforce its repressive policies toward women and women's rights in the country.[30]

Challenges

Dangerous, discriminatory, and hateful narratives about individuals and communities have existed for centuries and are unlikely to disappear. However, as technology dramatically assists with the spread and impact of these narratives, their persistence raises important questions about deliberative democracy, human security, and the quality of democratic listening in the digital era: Can deliberative democratic processes be distorted to the point where hate and discrimination drown out other voices, potentially leading to severe violence and societal harms? And how can we better leverage digital technologies to counter harmful narratives and build and strengthen a digital culture for peace?

Human Dimensions of Identity-Based Disinformation

Addressing these questions and the broader challenges posed by identity-based disinformation is a challenging problem because of

the human and technical dimensions. First, because identity-based disinformation feeds into an individual's preexisting biases, it can be more difficult to debunk or counter; a fact-check narrowly correcting a single rumor about an ethnic group makes only a minor dent in the larger narratives underpinning identity-based animosity. Unlike other forms of disinformation that rely on extreme or egregious falsehoods that contradict a person's existing beliefs, identity-based disinformation is often crafted to align with and exacerbate the prejudices that already exist within individuals and society. Thus, identity-based disinformation can play a social role within communities that perpetuate those prejudices, which can make it harder to debunk or mitigate, as it taps into deep-seated anxieties and resentments within the affected groups.

Relatedly, emotions such as fear and anxiety are often entwined into identity-based disinformation narratives. For example, research by the Migration Policy Institute highlights how fear, anger, and prejudice are used to perpetuate disinformation about migrants and migration.[31] For example, during the war in Ukraine, Russia has strategically co-opted preexisting anti-immigration sentiment to spread disinformation narratives about Ukrainian refugees as "criminals and rapists" who are "not welcome in Europe".[32] Ultimately, these strategies demonstrate how preexisting fears and anxieties surrounding immigration are manipulated to reinforce harmful stereotypes and exacerbate societal divisions.

Technological Dimensions of Identity-Based Disinformation

The challenges posed by identity-based disinformation also extend beyond human behavior; it is profoundly influenced by the design and implementation of technology. Platforms are built on models of surveillance capitalism and attention economics, which can exacerbate the spread of sensational or negative content.[33] Because identity-based disinformation exploits fracture lines within and across communities, it has the propensity to generate higher levels of engagement and be algorithmically enforced. Indeed, studies have demonstrated that highly emotional, false, and negative news content generates far more engagement than facts.[34] Thus, identity-based disinformation can be significantly shaped and propelled by technology and the algorithms and business models that prioritize sensational content and emotional engagement.

However, not all identity-based disinformation is fueled by algorithms; it can, for example, also circulate within closed networks of trusted communities on chat applications like WhatsApp. In these environments, disinformation can spread rapidly through personal connections in small communities characterized by shared beliefs, cultures, or heritages, making it difficult to challenge or fact-check. Indeed, researchers and journalists have highlighted evidence of mis- and disinformation spreading across closed platforms.[35] As internet users continue to shift to these platforms for political and interpersonal communication, thinking about their closed network features presents unique and ongoing opportunities, as well as challenges for combatting disinformation.[36]

At the same time, investment in trust and safety measures has been declining, while the need for localized contextual expertise is on the rise. Emerging scholarship on disinformation highlights the importance of the "cultural variation of speech acts, the normative orders bundled around them, and the historical conditions that underpin them".[37] However, nuanced evaluations of and responses to identity-based disinformation remain limited as platforms continue to underinvest in content monitoring and moderation in local languages. Without these investments, being able to detect and interrupt identity-based disinformation will be limited, since ethnic and racial slurs can often form swiftly through local insider language and from formerly innocuous language.

Finally, innovations in new and digital technologies are going to affect the challenges we are already facing today around combating identity-based disinformation. For instance, immersive video game platforms may turn into a new front for combatting the adoption of extremist identities and ideologies.[38] Additionally, generative AI technologies are increasingly being used to enhance targeted disinformation campaigns, particularly against female politicians and journalists, amplifying harassment and violence.[39] Thus, developing policies that are resilient and adaptable to the changing technological landscape will be important for creating a lasting impact.

Building a Digital Culture for Peace

Addressing identity-based disinformation will require more than just technological solutions; we must also focus on the human dimensions that can make certain narratives more salient. The ways in which people and algorithms interact and socially shape our digital environments play a crucial role in the spread and impact of identity-based disinformation. There are already many interventions aimed at combating the harms of disinformation, which are being deployed in different countries, on different platforms, and within different communities of users. Interventions include strategies such as fact-checking and correcting disinformation, improving the media literacy of users, supporting victims of disinformation and online harassment, and raising the costs of producing and distributing disinformation for political or economic gain. Considering the specific characteristics, impacts, and challenges of identity-based disinformation, what lessons can we draw from existing efforts and how might they be adapted or advanced further to mitigate harms?

Myanmar was one of the first high-profile cases of identity-based disinformation being weaponized to incite political violence, and as a case study it can offer some lessons for the broader global community. Following the high-profile revelations of the military-backed disinformation campaigns in the country, Facebook carried out a human rights impact assessment, which made several recommendations for improving the information environment and mitigating human rights abuses. Although this assessment came in 2018, after significant harm had already occurred, assessments nonetheless provided helpful recommendations for platform responses to

A Rohingya child watches from above as other children play chinlone, the national sport of Myanmar. (Salim Khan, Rohingya photographer/2018)

violence and human rights abuses, such as adopting a stand-alone human rights policy to oversee the company's human rights strategy, as well as recommending data preservation strategies that could be used in courts of law to evaluate international human rights violations.[40] Activists in Myanmar have suggested this could be a model not only for Myanmar but a process that is continually carried out in all markets Facebook operates in.[41]

Conducting a regular human rights impact assessment in all countries where platforms operate could help identify cases of identity-based disinformation early. For platforms maintaining civil society partnerships, these assessments could also offer better opportunities to (re)evaluate the impacts of their products in a way that is locally contextualized. Since 2018, Facebook has dramatically scaled up its content moderation efforts and engagement with civil society groups on the ground in Myanmar. Civil society actors have suggested that these efforts have gone a long way in promoting a more sophisticated understanding of the Myanmar context.[42] Further, human rights assessments can inform changes in platform design and interventions that can have downstream effects on behavior and outcomes, including deescalating conflicts long before violence takes place.[43] While there might not ever be a silver bullet solution to limiting the spread of identity-based disinformation and other forms of harmful content, ongoing collaboration, monitoring, and

resource sharing with local partners are some lessons learned from the Myanmar response that can help foster a safer digital environment for everyone.

Responding to identity-based disinformation will require a coherent strategy that can adapt to different cultural contexts as well as platform and media environments. In addition to the lessons learned from Myanmar, organizations are utilizing data-driven techniques to identify and analyze identity-based disinformation in other countries and have put forward promising new policy and programming recommendations.[44] Scholars and policymakers have highlighted and experimented with a variety of approaches to address the technical and societal dynamics that enable identity-based harms online. At the core of some of these recommendations is better content-moderation capacity and policy-violation enforcement by social media companies, with notable cautions about outsourcing content moderation to vulnerable local populations.[45] Additional moderation improvements could include better enforcement against accounts that call for violence,[46] or making technical design choices that mitigate harm, such as down-ranking divisive content by adding friction to the "angry" reaction on social media platforms.[47]

Additionally, strategies to combat identity-based disinformation harms could focus on improving reporting and safety mechanisms

for account holders,[48] creating helplines for individuals in high-profile and vulnerable positions, like female politicians,[49] or protecting female journalists who are covering sensitive political topics. Collaboration among different organizations and institutions—specifically platforms, institutions, and civil society organizations—will be important for creating a comprehensive approach to combating identity-based disinformation and ensuring the safety and security of individuals and communities online.

Responding to identity-based disinformation will require technical and human responses that are collaborative, locally relevant, and community driven. Given the extent of the harms posed by identity-based disinformation, there is a continued need to develop, implement, and improve responses. By continuing to allocate resources, share expertise, and foster partnerships, we can create a digital culture for peace that protects vulnerable individuals and communities from the harms of identity-based disinformation.

Bibliography

Afghan Witness. "Increase in Online Hate Speech Directed at Influential Afghan Women After Taliban Takeover." February 15, 2022. https://www.afghanwitness.org/reports/increase-in-online-hate-speech-directed-at-influential-afghan-women-since-taliban-takeover-.

Amnesty International. "The Social Atrocity: Meta and the right to remedy for the Rohingya." 2022, https://www.amnesty.org/en/documents/ASA16/5933/2022/en/.

Bail, Christopher A., Brian Guay, Emily Aloney, Aidan Combs, D. Sunshine Hillygus, Friedolin Merhout, Deen Freelon, and Alexander Volfovsky. "Assessing the Russian Internet Research Agency's Impact on the Political Attitudes and Behaviors of American Twitter Users in Late 2017." PNAS 117, no. 1 (November 25, 2019): 243-250.

Barrett, Paul M. "Fueling The Fire: How Social Media Intensifies U.S. Political Polarization – And What Can Be Done About It." NYU Stern Center for Business and Human Rights. 2021, https://bhr.stern.nyu.edu/publication/fueling-the-fire-how-social-media-intensifies-u-s-political-polarization-and-what-can-be-done-about-it/.

Barrett, Paul M. "'We Want You to Be a Proud Boy' How Social Media Facilitates Political Intimidation and Violence." NYU Stern Center for Business and Human Rights. September 2024, https://bhr.stern.nyu.edu/publication/how-social-media-facilitates-political-intimidation-and-violence/ .

Benesch, Susan. Countering Dangerous Speech: New Ideas for Genocide Prevention. Working paper. US Holocaust Memorial Museum. 2014. https://www.ushmm.org/m/pdfs/20140212-benesch-countering-dangerous-speech.pdf.

Bellovary et al. (2021). "Left-and-Right-Leaning News Organization Use Negative Emotional Content and Elicit User Engagement Similarly." PNAS. (2021):

Bradshaw et al (2020). Sourcing and Automation of Political News and Information over Social Media in the United States, 2016-2018. Political Communication.

Bradshaw and Henle (2021). The Gender Dimensions of Foreign Influence Operations. International Journal of Communication.

Bradshaw and Howard (2018). The Global Organization of Social Media Disinformation Campaigns. Journal of International Affairs 71 (1.5): 23-32.

Calderon, Maria. (2024). Political Violence in Mexico's 2024 Elections: Gender Based Political Violence Against Women. Wilson Center. Available: https://www.wilsoncenter.org/article/political-violence-mexicos-2024-elections-gender-based-political-violence-against-women.

Dangerous Speech Project (2021). Dangerous Speech: A Practical Guide. https://dangerousspeech.org/guide/.

Diamond, Larry. 2019. The Road to Digital Unfreedom: The Threat of Post-modern Totalitarianism. Journal of Democracy, 30(1):23-24.

Deibert, Ron. (2019). Three Painful Truths about Social Media. Journal of Democracy, 30(1):25-39.

Druckman, James N. et al, "Affective polarization, local contexts and public opinion in America," Nature Human Behaviour 5, no. 28-38 (2021).

European Union High Level Working Group on Fake News and Online Disinformation. (2018). https://digital-strategy.ec.europa.eu/en/library/final-report-high-level-expert-group-fake-news-and-online-disinformation.

Feldstein, Steven. (2021). The Rise of Digital Repression: How Technology Is Reshaping Power, Politics and Resistance.

Freelon, Deen et al. "Black Trolls Matter: Racial and Ideological Asymmetries in Social Media Disinformation." 2020, https://journals.sagepub.com/doi/abs/10.1177/0894439320914853.

Global Engagement Center. (2022). PRC Efforts to Manipulate Global Public Opinion on Xinjiang. Global Engagement Center. Available: https://www.state.gov/prc-efforts-to-manipulate-global-public-opinion-on-xinjiang/.

Global Witness (2022). Facebook approves adverts containing hate speech inciting violence and genocide

against the Rohingya. *Global Witness*. Available: https://www.globalwitness.org/en/campaigns/digital-threats/rohingya-facebook-hate-speech/.

Grosz, Elizabeth. (1990) A note on essentialism and difference. In: Gunew S(ed.) *Feminist Knowledge: Critique and Construct*. London: Routledge.

Hao, Karen. (2021). The Facebook Whistleblower says its Algorithms are Dangerous. Here's why. *MIT Technology Review*. Available: https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/.

Heller (2021). Enlisting Useful Idiots: The Ties Between Online Harassment and Disinformation. *Colorado Technology Law Journal*.

Hola, Barbora, Hollie Nyseth Nzitatira, and Maartje Weerdesteijn. (2022). *Introduction: Atrocity Crimes and Atrocity Studies*. Oxford University Press. Oxford: UK.

Hook, Kristina and Ernesto Verdeja. (2022). Social Media Misinformation and the Prevention of Political Instability and Mass Atrocities. *Stimson Center*. Available: https://www.stimson.org/2022/social-media-misinformation-and-the-prevention-of-political-instability-and-mass-atrocities/.

Howard, Philip N. et al, "The IRA, Social Media and Political Polarization in the United States, 2012-2018." University of Oxford, 2018.

Hynes (2017). Internet Use is on the rise in Myanmar, but better options are needed. https://www.forbes.com/sites/chynes/2017/09/22/internet-use-is-on-the-rise-in-myanmar-but-better-options-are-needed/?sh=56c3fc10448e.

International Press Institute. 2016. Recommendations on protecting journalists from online harassment. *International Press Institute, Available:* https://ipi.media/recommendations-on-protecting-journalists-bloggers-from-online-harassment/.

Jack (2017). Lexicon of Lies. *Data and Society.* https://datasociety.net/library/lexicon-of-lies/.

Jackson, Dean and Jon Bateman. 2024. Countering Disinformation Effectively: AN Evidence-Based Policy Guide. *Carnegie Endowment for International Peace*. Available: https://carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide?lang=en.

Jancowicz et al. (2021). Malign Creativity: How Gender, Sex and Lies are Weaponized Against Women Online. *The Wilson Center.* https://www.wilsoncenter.org/publication/malign-creativity-how-gender-sex-and-lies-are-weaponized-against-women-online.

Jensen, Michael. (2018). Russian Trolls and Fake News: Information or Identity Logics. *Journal of International Affairs*, 71(15).

Klepper, David. (2024). Sexist Tropes and Misinformation Swirl Online as Mexico Prepares to Elect its first Female Leader. *AP News*, Available: https://apnews.com/article/mexico-election-sheinbaum-facebook-lopez-obrador-79adddaf-8300f30af51fddbbf3165216.

Lakhani, Suraj. (2021). Video Gaming and (Violent) Extremism: An Exploration of the Current Landscape, Trends and Threats. *European Commission.*

Larsen, Anna Grondahl, Ingrid Fadnes and Roy Krovel. 2020. *Journalist Safety and Self -Censorship.* Routledge. Available: https://doi.org/10.4324/9780367810139.

Long (2018). Why Facebook is cracking down on Myanmar Military Accounts. https://time.com/5383780/myanmar-facebook-propaganda-rohingya/.

Lukito, Josephine et al. "The Wolves in Sheep's Clothing: How Russia's Internet Research Agency Tweets Appeared in U.S. News as Vox Populi." The International Journal of Press/Politics. 2019, https://journals.sagepub.com/doi/abs/10.1177/1940161219895215?journalCode=hijb.

Malhorta, Pranav. (2023). Misinformation in WhatsApp Family Groups: Generational Perceptions and Correction Considerations in a Meso-News Space. *Digital Journalism*, 12(5). https://doi.org/10.1080/21670811.2023.2213731.

Monaco, Nicholas and Carly Nyst. "State-Sponsored Trolling: How Governments Are Deploying Disinformation as Part of Broader Digital harassment Campaigns." Institute for the Future. 2018, https://legacy.iftf.org/statesponsoredtrolling/.

Meta. "Coordinated Inauthentic Behavior," December 6, 2018, https://about.fb.com/news/tag/coordinated-inauthentic-behavior/.

Meta Newsroom. (2018). Removing Myanmar Military Officials from Facebook. https://about.fb.com/news/2018/08/removing-myanmar-officials/.

Marais, Michael. "How disinformation is increasingly weaponised against women and minority groups." Centre for Information Resilience, 2024, https://www.info-res.org/post/how-disinformation-is-increasingly-weaponised-against-women-and-minority-groups.

Markham, Susan and Moira Whelan. (2024). Rapid Online Support or Women in Politics. *National Democratic Institution*. Available: https://www.ndi.org/our-stories/rapid-online-support-women-politics.

Mimizuka, Kayo and Inga Trauthig. (2022). WhatsApp Misinformation and Latino Political Discourse in the U.S. *Tech Policy Press*. Available: https://www.techpolicy.press/whatsapp-misinformation-and-latino-political-discourse-in-the-u-s/.

Mozur (2018). A Genocide incited on Facebook, With Posts from Myanmar's Military. https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html.

Neidhardt, Alberto-Horst and Paul Butcher. (2022). Disinformation on Migration: How Lies, Half-Truths and Mischaracterizations Spread. *Migration Policy Institute.* Available: https://www.migrationpolicy.org/article/how-disinformation-fake-news-migration-spreads.

Ozawa, Joao V. S., Samuel C. Woolley, Joseph Straubhaar, Martin J. Riedl, Katie Joseff and Jacob Gursky. (2023). How Disinformation on WhatsApp went from Campaign Weapon to Government Propaganda in Brazil. S*ocial Media + Society*, 9(1). https://doi.org/10.1177/20563051231160632.

Posetti, Julie and Nabeelah Shabbir. (2022). The Chilling: A Global Study of Online Violence Against Women Journalists. *International Center for Journalists.* Available: https://www.icfj.org/our-work/chilling-global-study-online-violence-against-women-journalists.

Reddi, Kuo and Kreiss (2021). Identity Propaganda: Racial Narratives and Disinformation. *New Media and Society.*

Rio, Victorie. (2021). "Myanmar: The Role of Social Media in Fomenting Violence." In Schirch, Lisa (Ed) *Social Media Impacts on Conflict and Democracy: The Techtonic Shift.* Routledge Advances in International Relations and Global Politics.

Rosenblat, Mariana Olaizola. (2023). Gaming the System: How Extremists Exploit Gaming Sites and What Can Be Done to Counter them. *NYU Stern Center for Business and Human Rights.* Available: https://bhr.stern.nyu.edu/publication/gaming-the-system-how-extremists-exploit-gaming-sites-and-what-can-be-done-to-counter-them/.

Sarrats, Mireia Faro. (2024). Fear and Fatigue: Why Anti-Migrant Sentiment in Europe Helps Russia. *European Council on Foreign Relations.* Available: https://ecfr.eu/article/fear-and-fatigue-why-anti-migrant-sentiment-in-europe-helps-russia/.

Scott, Mark. (2024). Deepfake Porn is Political Violence. *Politico.* Available: https://www.politico.eu/newsletter/digital-bridge/deepfake-porn-is-political-violence/.

Shah, Sayyed Fawad Ali, Ivana Cvetkovic, Tamar Ginossar, Rahman Ullah, Danish Barber and Autumn Slaughter. 2024. Online Harassment, Psychological Stressors, and Occupational Dysfunction among Journalists working in a conflict zone. *Digital Journalism*, 12 (6): 735-752.

Stecklow (2018). Why Facebook is Losing the War on Hate Speech in Myanmar. https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

Stray, Jonathan, Ravi Iyer and Helena Puig Larrauri. (2023). The Algorithmic Management of Polarization and Violence on Social Media. *Knight First Amendment Institute, Columbia University.* Available: https://knightcolumbia.org/content/the-algorithmic-management-of-polarization-and-violence-on-social-media.

Udupa et al. (2021) Extreme Speech in Global Digital Cultures. *International Journal of Communication.*

UNESCO (2020). Online Violence Against Women Journalists: A Global Snapshot of Incidence and Impacts. Available: https://unesdoc.unesco.org/ark:/48223/pf0000375136.

Vraga and Bode (2020). Defining Misinformation and Understanding its Bounded Nature: Using Expertise and Evidence for Describing Misinformation. *Political Communication.*

Ward, Megan and Jessica Beyer. (2019). Vulnerable Landscapes: Case Studies of Violence and Disinformation. *Wilson Center.* Available: https://www.wilsoncenter.org/publication/vulnerable-landscapes-case-studies-violence-and-disinformation.

Wardle, Claire. (2017). Understanding the Information Disorder. *First Draft News.*

Warofka, Alex. (2018). An Independent Assessment of the Human Rights Impact of Facebook in Myanmar. *Meta Newsroom,* November 5. Available: https://about.fb.com/news/2018/11/myanmar-hria/.

Wu, Tim. (2016). *The Attention Merchants: The Epic Scramble to Get Inside our Heads.*

Yung Yao, Tsai and Jonathan Chin. 2024. China is Posting Fake Videos of President. *Taipei Times,* Available: https://www.taipei-times.com/News/front/archives/2024/01/11/2003811930.

Zuboff, Shoshana. (2018). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. *Public Affairs: New York.*

Endnotes

1   Paul Mozur, "A Genocide incited on Facebook, With Posts from Myanmar's Military," *The New York Times*, 2018, https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html.

2   Kayleigh Long, "Why Facebook is cracking down on Myanmar Military Accounts," *TIME*, 2018, https://time.com/5383780/myanmar-facebook-propaganda-rohingya/.

3   Steve Stecklow, "Why Facebook is Losing the War on Hate Speech in Myanmar," *Reuters*, 2018, https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

4   Casey Hynes, "Internet Use is on the rise in Myanmar, but better options are needed," *Forbes*, 2017, https://www.forbes.com/sites/chynes/2017/09/22/internet-use-is-on-the-rise-in-myanmar-but-better-options-are-needed/?sh=56c3fc10448e.

5   Amnesty International, "The Social Atrocity: Meta and the right to remedy for the Rohingya," 2022, https://www.amnesty.org/en/documents/ASA16/5933/2022/en/.

6   Alex Warofka, "An Independent Assessment of the Human Rights Impact of Facebook in Myanmar," *Meta Newsroom*, November 5, 2018, https://about.fb.com/news/2018/11/myanmar-hria/; "Removing Myanmar Military Officials from Facebook," Meta Newsroom, 2022, https://about.fb.com/news/2018/08/removing-myanmar-officials/.

7   "Facebook approves adverts containing hate speech inciting violence and genocide against the Rohingya," *Global Witness*, 2022, https://www.globalwitness.org/en/campaigns/digital-threats/rohingya-facebook-hate-speech/.

8   Larry Diamond, "The Road to Digital Unfreedom: The Threat of Post-modern Totalitarianism," *Journal of Democracy* 30, no. 1 (2019): 23-24; Ron Deibert, "Three Painful Truths about Social Media," *Journal of Democracy* 30, no. 1 (2019): 25-39; Samantha Bradshaw and Philip N. Howard, "The Global Organization of Social Media Disinformation Campaigns," *Journal of International Affairs* 71, no. 1.5: 23-32; Steven Feldstein, *The Rise of Digital Repression: How Technology Is Reshaping Power, Politics and Resistance* (2021).

9   Christopher A. Bail et al, "Assessing the Russian Internet Research Agency's Impact on the Political Attitudes and Behaviors of American Twitter Users in Late 2017," PNAS 117, no. 1 (2019): 243-250; Philip N. Howard et al, "The IRA, Social Media and Political Polarization in the United States, 2012-2018," University of Oxford, 2018; Josephine Lukito et al, "The Wolves in Sheep's Clothing: How Russia's Internet Research Agency Tweets Appeared in U.S. News as Vox Populi", The International Journal of Press/Politics, 2019, https://journals.sagepub.com/doi/abs/10.1177/1940161219895215?journalCode=hijb.

10 European Union High Level Expert Group on Fake News and Online Disinformation, "A multi-dimensional approach to disinformation," 2018, https://digital-strategy.ec.europa.eu/en/library/final-report-high-level-expert-group-fake-news-and-online-disinformation; Caroline Jack, "Lexicon of Lies," *Data and Society*, 2017, https://datasociety.net/library/lexicon-of-lies/; Claire Wardle, "Understanding the Information Disorder," *First Draft News*, 2017; Emily K. Vraga and Leticia Bode, "Defining Misinformation and Understanding its Bounded Nature: Using Expertise and Evidence for Describing Misinformation," *Political Communication*, 2020.

11 "Coordinated Inauthentic Behavior," Meta, December 6, 2018, https://about.fb.com/news/tag/coordinated-inauthentic-behavior/.

12 Brittan Heller, "Enlisting Useful Idiots: The Ties Between Online Harassment and Disinformation," *Colorado Technology Law Journal*, 2021.

13 Samantha Bradshaw and Amélie Henle, "The Gender Dimensions of Foreign Influence Operations," *International Journal of Communication*; Madhavi Reddi et al, "Identity Propaganda: Racial Narratives and Disinformation," *New Media and Society*, 2021; Nina Jankowicz et all, "Malign Creativity: How Gender, Sex and Lies are Weaponized Against Women Online," *The Wilson Center*, https://www.wilsoncenter.org/publication/malign-creativity-how-gender-sex-and-lies-are-weaponized-against-women-online.

14 Susan Benesch, "Countering Dangerous Speech: New Ideas for Genocide Prevention," United States Holocaust Memorial Museum, 2014, https://www.ushmm.org/m/pdfs/20140212-benesch-countering-dangerous-speech.pdf.

15 Dangerous Speech Project, "Dangerous Speech: A Practical Guide," 2021, https://dangerousspeech.org/guide/.

16 Michael Jensen, "Russian Trolls and Fake News: Information or Identity Logics," *Journal of International Affairs* 71, no. 15 (2018); Madhavi Reddi et al, "Identity Propaganda;" Samantha Bradshaw and Amélie Henle, "Gender Dimensions;" Deen Freelon et al, "Black Trolls Matter: Racial and Ideological Asymmetries in Social Media Disinformation," 2020, https://journals.sagepub.com/doi/abs/10.1177/0894439320914853.

17 Madhavi Reddi et al, "Identity Propaganda;" Elizabeth Grosz, "A note on essentialism and difference," in *Feminist Knowledge: Critique and Construct*, ed. Sneja Gunew (Routledge, 1990).

18   James N. Druckman et al, "Affective polarization, local contexts and public opinion in America," *Nature Human Behaviour* 5, no. 28-38 (2021); Paul M. Barrett, "Fueling The Fire: How Social Media Intensifies U.S. Political Polarization – And What Can Be Done About It," NYU Stern Center for Business and Human Rights, 2021, https://bhr.stern.nyu.edu/publication/fueling-the-fire-how-social-media-intensifies-u-s-political-polarization-and-what-can-be-done-about-it/.

19   Samantha Bradshaw and Amélie Henle, "Gender Dimensions."

20   María Calderón, "Political Violence in Mexico's 2024 Elections: Gender Based Political Violence Against Women," *Wilson Center*, https://www.wilsoncenter.org/article/political-violence-mexicos-2024-elections-gender-based-political-violence-against-women.

21   Nicholas Monaco and Carly Nyst, "State-Sponsored Trolling: How Governments Are Depoloying Disinformation as Part of Broader Digital harassment Campaigns," Institute for the Future, 2018, https://legacy.iftf.org/statesponsoredtrolling/.

22   Julie Posetti et al, "Online Violence Against Women Journalists: A Global Snapshot of Incidence and Impacts," UNESCO, 2020, https://unesdoc.unesco.org/ark:/48223/pf0000375136.

23   Julie Posetti and Nabeelah Shabbir, "The Chilling: A Global Study of Online Violence Against Women Journalists," *International Center for Journalists*, 2022, https://www.icfj.org/our-work/chilling-global-study-online-violence-against-women-journalists.

24   Sayyed Fawad Ali Shah et al, "Online Harassment, Psychological Stressors, and Occupational Dysfunction among Journalists Working in a Conflict Zone," *Digital Journalism* 12, no. 6 (2024): 735-752; Anna Grøndahl Larsen et al, eds., *Journalist Safety and Self-Censorship* (Routledge, 2020), https://doi.org/10.4324/9780367810139.

25   Kayo Mimizuka and Inga Trauthig, "WhatsApp Misinformation and Latino Political Discourse in the U.S.," *Tech Policy Press*, 2022, https://www.techpolicy.press/whatsapp-misinformation-and-latino-political-discourse-in-the-u-s/.

26   Barbora Hola et al, *Introduction: Atrocity Crimes and Atrocity Studies* (Oxford University Press, 2022).

27   Megan Ward and Jessica Beyer, "Vulnerable Landscapes: Case Studies of Violence and Disinformation," *Wilson Center*, 2019, https://www.wilsoncenter.org/publication/vulnerable-landscapes-case-studies-violence-and-disinformation.

28   "PRC Efforts to Manipulate Global Public Opinion on Xinjiang," *Global Engagement Center*, 2022, https://www.state.gov/prc-efforts-to-manipulate-global-public-opinion-on-xinjiang/.

29   Karen Hao, "The Facebook Whistleblower says its Algorithms are Dangerous. Here's why," *MIT Technology Review*, 2021, https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/.

30   "Increase in Online Hate Speech Directed at Influential Afghan Women After Taliban Takeover," *Afghan Witness*, February 15, 2022, https://www.afghanwitness.org/reports/increase-in-online-hate-speech-directed-at-influential-afghan-women-since-taliban-takeover-.

31   Alberto-Horst Neidhardt and Paul Butcher, "Disinformation on Migration: How Lies, Half-Truths and Mischaracterizations Spread," *Migration Policy Institute*, 2022, https://www.migrationpolicy.org/article/how-disinformation-fake-news-migration-spreads.

32   Mireia Faro Sarrats, "Fear and Fatigue: Why Anti-Migrant Sentiment in Europe Helps Russia," *European Council on Foreign Relations*, 2024, https://ecfr.eu/article/fear-and-fatigue-why-anti-migrant-sentiment-in-europe-helps-russia/.

33   Shoshana Zuboff, "The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power," *Public Affairs: New York*, 2018; Tim Wu, *The Attention Merchants: The Epic Scramble to Get Inside our Heads*, 2018.

34   Samantha Bradshaw et al, "Sourcing and Automation of Political News and Information over Social Media in the United States, 2016-2018," *Political Communication*, 2020; Andrea K. Bellovary et al, "Left-and-Right-Leaning News Organization Use Negative Emotional Content and Elicit User Engagement Similarly," PNAS (2021).

35   Joao V. S. Ozawa et al, "How Disinformation on WhatsApp went from Campaign Weapon to Government Propaganda in Brazil," *Social Media + Society* 9, no. 1 (2023), https://doi.org/10.1177/20563051231160632.

36   Pranav Malhorta, "Misinformation in WhatsApp Family Groups: Generational Perceptions and Correction Considerations in a Meso-News Space," *Digital Journalism* 12, no. 5 (2023), https://doi.org/10.1080/21670811.2023.2213731.

37   Sahana Udupa and Matti Pohjonen, "Extreme Speech and Global Digital Cultures," *International Journal of Communication*, 2019.

38 Suraj Lakhani, "Video Gaming and (Violent) Extremism: An Exploration of the Current Landscape, Trends and Threats," *European Commission*, 2021; Mariana Olaizola Rosenblat, "Gaming the System: How Extremists Exploit Gaming Sites and What Can Be Done to Counter Them," NYU *Stern Center for Business and Human Rights*, 2023, https://bhr.stern.nyu.edu/publication/gaming-the-system-how-extremists-exploit-gaming-sites-and-what-can-be-done-to-counter-them/.

39 Mark Scott, "Deepfake Porn is Political Violence," *Politico*, 2024, https://www.politico.eu/newsletter/digital-bridge/deepfake-porn-is-political-violence/; David Klepper, "Sexist Tropes and Misinformation Swirl Online as Mexico Prepares to Elect its first Female Leader," *AP News*, 2024, https://apnews.com/article/mexico-election-sheinbaum-facebook-lopez-obrador-79adddaf8300f30af51fddbbf3165216; Tsai Yung Yao and Jonathan Chin, "China is Posting Fake Videos of President," *Taipei Times*, 2024, https://www.taipeitimes.com/News/front/archives/2024/01/11/2003811930.

40 Alex Warofka, "Human Rights Impact."

41 Victorie Rio, "Myanmar: The Role of Social Media in Fomenting Violence," in *Social Media Impacts on Conflict and Democracy: The Techtonic Shift*, ed. Lisa Schirch (Routledge, 2021).

42 Victorie Rio, "Fomenting Violence."

43 Jonathan Stray et al, "The Algorithmic Management of Polarization and Violence on Social Media," *Knight First Amendment Institute, Columbia University*, 2023, https://knightcolumbia.org/content/the-algorithmic-management-of-polarization-and-violence-on-social-media.

44 Michael Marais, "How disinformation is increasingly weaponised against women and minority groups," Centre for Information Resilience, 2024, https://www.info-res.org/post/how-disinformation-is-increasingly-weaponised-against-women-and-minority-groups.

45 Paul M. Barrett, "'We Want You to Be a Proud Boy' How Social Media Facilitates Political Intimidation and Violence," NYU Stern Center for Business and Human Rights, September 2024, https://bhr.stern.nyu.edu/publication/how-social-media-facilitates-political-intimidation-and-violence/

46 Kristen Hook and Ernesto Verdeja, "Social Media Misinformation and the Prevention of Political Instability and Mass Atrocities," *Stimson Center*, 2022, https://www.stimson.org/2022/social-media-misinformation-and-the-prevention-of-political-instability-and-mass-atrocities/.

47 Paul M. Barrett, "Proud Boy;" Dean Jackson and Jon Bateman, "Countering Disinformation Effectively: AN Evidence-Based Policy Guide," *Carnegie Endowment for International Peace*, 2024, https://carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide?lang=en; Stray et al, "Algorithmic Management."

48 "Recommendations on protecting journalists from online harassment," *International Press Institute*, 2016, https://ipi.media/recommendations-on-protecting-journalists-bloggers-from-online-harassment/.

49 Susan Markham and Moira Whelan, "Rapid Online Support or Women in Politics," *National Democratic Institution*, 2024, https://www.ndi.org/our-stories/rapid-online-support-women-politics.

About the Author

Dr. Samantha Bradshaw is the Director of the Center for Security, Innovation and New Technology at American University, and an Assistant Professor in the Department of Foreign Policy & Global Security in the School of International Service. Her research and public writing examine issues around technology, democracy, and security, including disinformation and foreign influence operations. As a leading expert on new technologies and democracy, Dr. Bradshaw's research has been featured in outlets such as the *New York Times*, *Washington Post*, *CNN* and *Bloomberg Magazine*, and she has published in several academic journals, including New Media & Society, Political Communication, and the International Journal of Communication. Dr. Bradshaw obtained her D.Phil. from Oxford University in 2020 and completed a postdoctoral fellowship at Stanford University between 2020-2022.

Cover photo: A young Rohingya girl poses behind the door of her shelter in a refugee camp in Bangladesh. (Sahat Zia Hero/2023)

About Us

The Stanley Center for Peace and Security partners with people, organizations, and the greater global community to drive policy progress in three issue areas—mitigating climate change, avoiding the use of nuclear weapons, and preventing mass violence and atrocities. The center was created in 1956 and maintains its independence while developing forums for diverse perspectives and ideas. To learn more about our recent publications and upcoming events, please visit stanleycenter.org.

About the Center for Security, Innovation, and Technology

The Center for Security, Innovation, and New Technology is an interdisciplinary program applying cutting edge research to security, power, and policy dilemmas stemming from digital and emerging technologies. Based at the School of International Service at American University, our faculty and students work on issues such as interstate cyber conflict, the proliferation and mitigation of spyware, disinformation and foreign influence operations, and the geopolitics of frontier technologies.

11/24