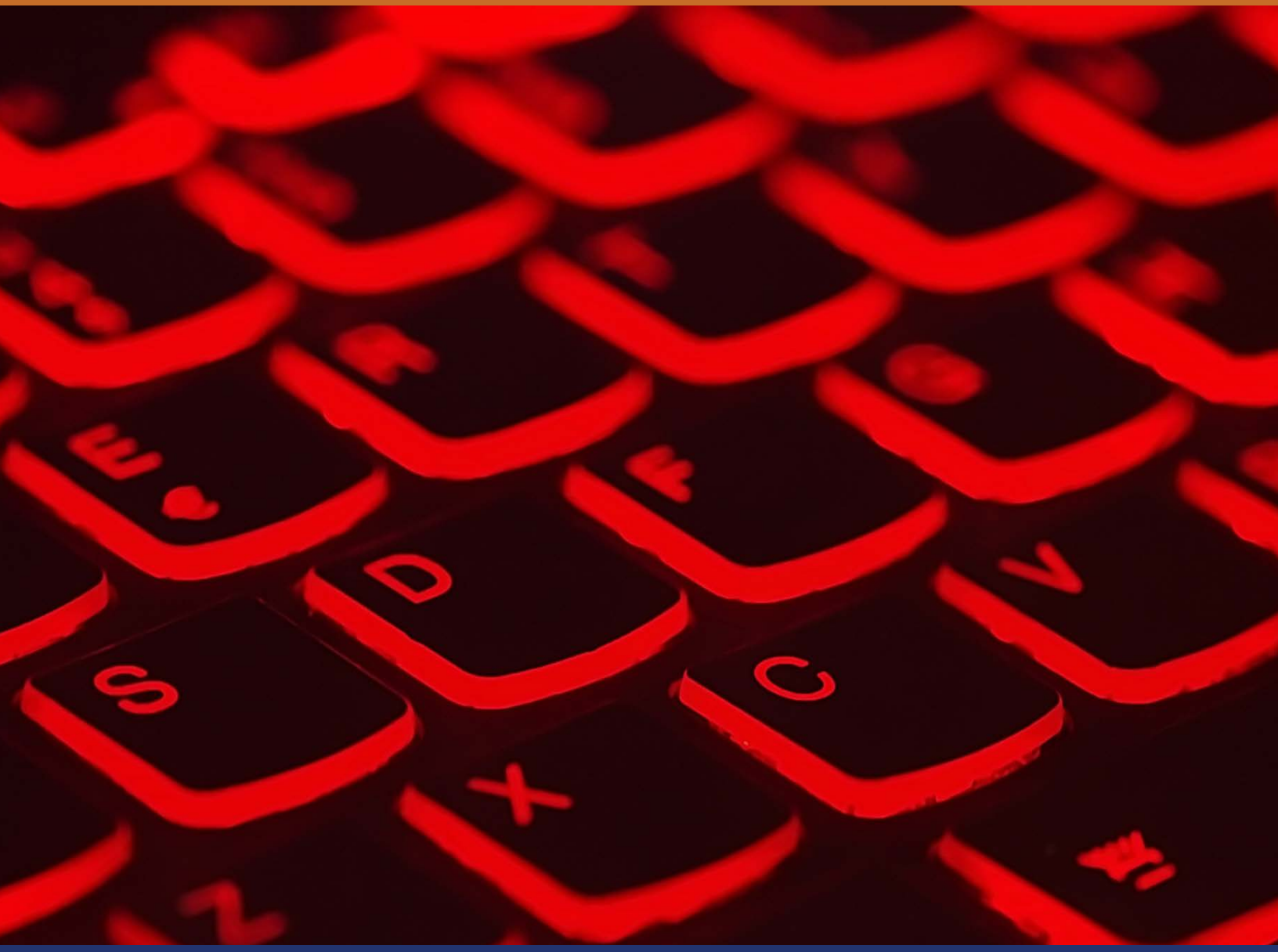
Disinformation Campaigns and Hate Speech: Exploring the Relationship and Programming Interventions

April 2019



INTERNATIONAL FOUNDATION FOR ELECTORAL SYSTEMS
IFES
30 Years

Disinformation Campaigns and Hate Speech:

Exploring the Relationship and Programming Interventions

April 2019

Lisa Reppell and Erica Shein
International Foundation for Electoral Systems

The authors would also like to thank Vasu Mohan, Chad Vickery, Katherine Ellena, Dr. Gabrielle Bardall and Heather Szilagyi for their invaluable reviews and perspectives. We are also grateful to Keaton Van Beveren for her graphic design.

International Foundation for Electoral Systems
2011 Crystal Drive, 10th Floor
Arlington, VA 22202
Email: editor@ifes.org
Fax: 202.350.6701

Introduction and Purpose of this Brief

Malign actors are increasingly deploying technology-fueled disinformation campaigns – rife with widely shared, inaccurate and polarizing information – around the globe. These campaigns amplify deep-seated sources of tension, discord and hatred in ways that undermine public trust in democratic institutions and increase the possibility of electoral violence and political instability. This brief defines the contours of this new generation of disinformation campaigns in relation to the scourge of hate speech. Then, it elaborates on the ways in which these campaigns, in their manipulation and exploitation of a changing media and information environment, increase and intensify hate speech already circulating in a political or electoral context.

The calculated amplification of hate speech is only one of many tactics deployed in disinformation campaigns, but it is a common and highly toxic one worthy of particular attention. This document focuses on the relationship between hate speech and disinformation and the implications of that relationship for the design of interventions to counter these dual threats. Specifically, we ask three questions:

- How is hate speech created and spread (its *trajectory*), independent of a technology-fueled disinformation campaign?
- How does that trajectory change with the introduction of a disinformation campaign?
- How does this altered trajectory impact the way that democracy and governance practitioners develop and deploy program interventions?

A strong body of research and practitioner resources provide guidance on how to address and mitigate hate speech that spreads organically through personal and virtual networks. The International Foundation for Electoral Systems (IFES) white paper *Countering Hate Speech in Elections: Strategies for Electoral Management Bodies*[1] draws from IFES' experience countering hate speech to identify best practices in this area. The paper identifies a comprehensive range of interventions against hate speech, including programming to promote tolerance, shift public conversations, monitor and report occurrences of hate speech and pursue appropriate legal remedies. Experience with these techniques provides an essential foundation upon which to further develop programming that responds to the spread of hate speech in a continually changing new media environment. For example, programs focusing on changing individual behaviors or beliefs can be combined with additional interventions elsewhere along the "chain of harm" to address hate speech that is generated and amplified in a technology-fueled disinformation campaign. As will be described further below, the goal of programming in these contexts should be to disrupt this chain of harm at multiple points, which include the *actor, message, mode of dissemination, interpreter*, and *risk.*

[1] Mohan, Vasu and Catherine Barnes (2018).

Terms and Definitions

Hate speech refers to polarizing expression that vilifies, humiliates or promotes intolerance and violence against groups of persons by explicit or indirect reference to their race, national or ethnic origin, religion, gender, sexual orientation, age, disability or other shared identity.[2] The negative impacts of hate speech on democracy can be considerable, especially the harms posed to the specific groups targeted by such expression. A 2014 ruling of the Supreme Court of India (Sangathan v. Union of India) is instructive in this regard:

> *Hate speech is an effort to marginalize individuals based on their membership in a group. Using expression that exposes the group to hatred, hate speech seeks to delegitimise group members in the eyes of the majority, reducing their social standing and acceptance within society. Hate speech, therefore, rises beyond causing distress to individual group members. It can have a societal impact. Hate speech lays the groundwork for later, broad attacks on [the] vulnerable that can range from discrimination, to ostracism, segregation, deportation, violence and, in the most extreme cases, to genocide. Hate speech also impacts a protected group's ability to respond to the substantive ideas under debate, thereby placing a serious barrier to their full participation in our democracy.*

For additional exploration of human rights and legal definitions of hate speech and how hate speech is particularly damaging to the groups it targets, consult IFES' *Countering Hate Speech in Elections* white paper.

For the purposes of this document, the phrase "disinformation problem" or "disinformation campaign" refers to the actions of inauthentic actors, either coordinated or lone, using technological means to produce or artificially amplify disinformation and malinformation, as defined below.[3] Additional definitions are also provided in the glossary at the end of this brief.

---

[2] Despite its ubiquity and virulence, there is no single, consensus definition for this concept among scholars, practitioners, and legal drafters. Definitions vary greatly depending on whether the creator is seeking a theoretical or academic understanding of the concept, or trying to intervene against it. This distinction is made in a recent scholarly effort to define the concept of hate speech, published by the Berkman Klein Center for Internet and Society at Harvard University. The report finds that "Where other areas of content analysis have developed rich methodologies to account for influences like context or bias, the present scholarship around hate speech rarely extends beyond identification of particular words or phrases that are likely to cause harm targeted toward immutable characteristics." https://cyber.harvard.edu/publications/2016/DefiningHateSpeech
Some definitions and sources for understanding the hate speech concept in all its permutations are available in the IFES paper by Mohan, Vasu and Catherine Barnes (2018).

[3] The definitions we have used herein are adapted from understandings of dis- mis- and mal-information around which the academic and practitioner communities have begun to coalesce. Similarly, our notion of inauthentic actors has been adapted from Facebook's definition of "coordinated inauthentic behavior," as it offers a useful way to distinguish between hate speech that spreads organically through personal and virtual networks, and hate speech that is generated or amplified as a tactic in a technology-fueled disinformation campaign.

Disinformation is false or misleading information that is created or disseminated with the intent to cause harm or to benefit the perpetrator. The intent to cause harm may be directed toward individuals, groups, institutions or processes.

Malinformation is accurate information that is shared with the intent to cause harm or to benefit the perpetrator, often by moving private information into the public sphere.

Misinformation is false or misleading information that is shared without the intent to cause harm or realization that it is incorrect. In some cases, actors may unknowingly perpetuate the spread of disinformation by sharing content they believe to be accurate among their networks.

Inauthentic actors are individuals or organizations working to mislead others about who they are or what they are doing.

Disinformation is not a new phenomenon. The sharing of false and misleading content is an age-old political tactic. Advances in technology and changes to the media environment are what set the contemporary "disinformation problem" apart. Similar to hate speech, the "disinformation problem" has a multitude of conceptual frames. It has been characterized as information disorder,[4] information manipulation,[5] information war,[6] computational propaganda[7] and coordinated inauthentic behavior.[8] What these frames share in common is an understanding that disinformation in today's new media environment is not only a problem of false information; it is a distortion of and an attack on our entire information ecosystem. It makes the free exchange of ideas much more difficult, if not impossible, in some contexts, undermining an essential element of any functioning democracy.

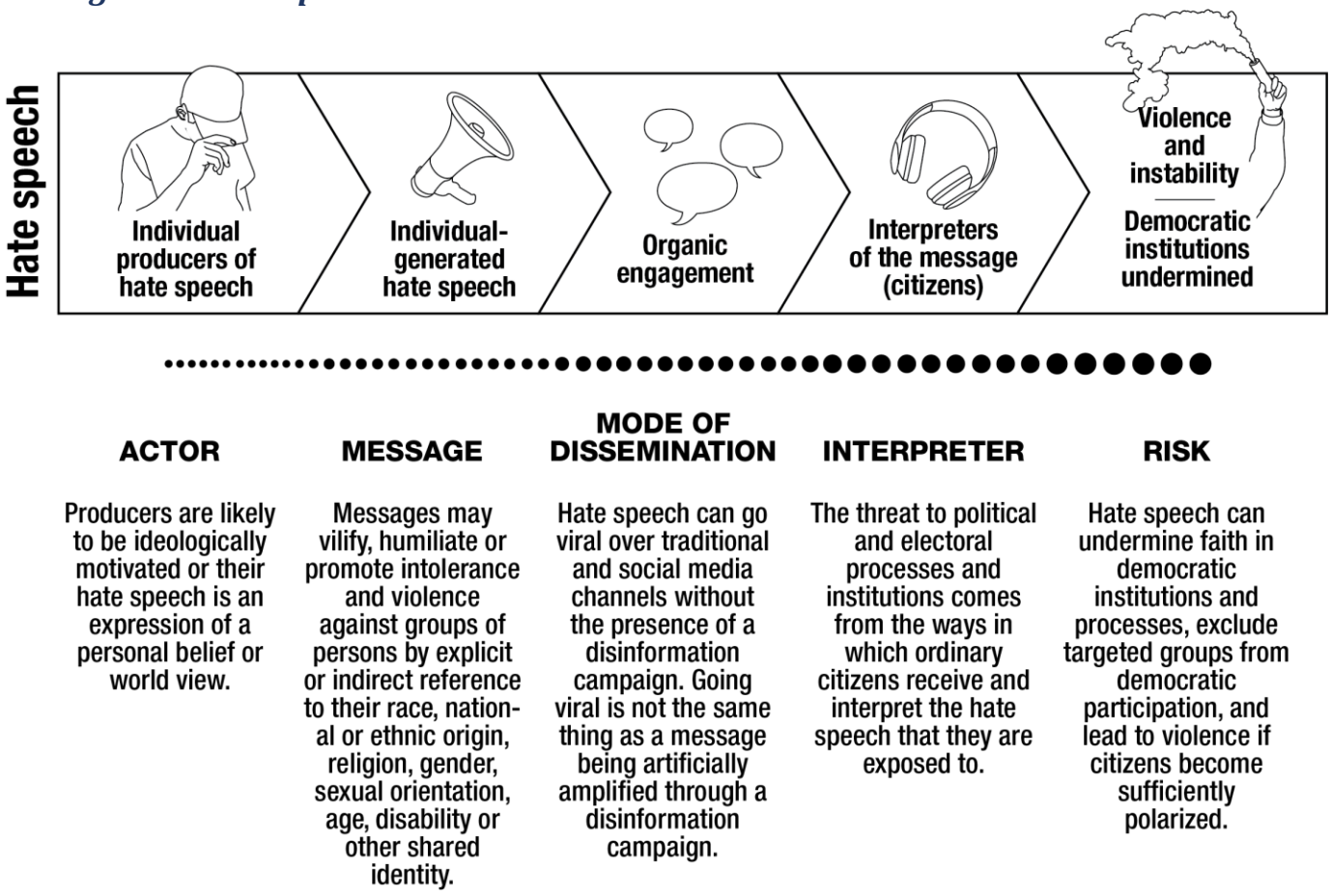
[4] Wardle, C. (2017). "Information Disorder: Toward an interdisciplinary framework for research and policy making," Council of Europe.
[5] Jeangène Vilmer, Jean-Baptiste, et al. (2018). "Information Manipulation," Center for Analysis, Prevision and Strategy of the Ministry for Europe and Foreign Affairs/Institute for Strategic Research of the Ministry for the Armed Forces (CAPS/IRSEM).
[6] DiResta, R. (2018). Statement for the record to the United States Senate Select Committee on Intelligence.
[7] Oxford Internet Institute
[8] Facebook uses the concept of "coordinated inauthentic behavior" to describe problematic networks of pages and people on their platform that "work together to mislead others about who they are or what they're doing." "Coordinated Inauthentic Behavior Explained," (2018). Facebook Newsroom.

Understanding the Relationship Between Hate Speech and Disinformation Campaigns

Disinformation campaigns that use hate speech as a tactic rely and build on underlying social dynamics and existing divisive messages and affinity groups. As noted above, there is a wide-ranging body of literature and practitioner resources focused on how hate speech is generated and disseminated and the impacts it has on the groups it targets. As the purpose of this brief is to situate hate speech within the new generation of technology-fueled disinformation campaigns, we focus in this section on understanding the trajectory of hate speech in elections, with and without the presence of such campaigns.

The first graphic below displays the typical trajectory of hate speech in the electoral context, absent technology-fueled disinformation campaigns.

Figure 1: Hate Speech



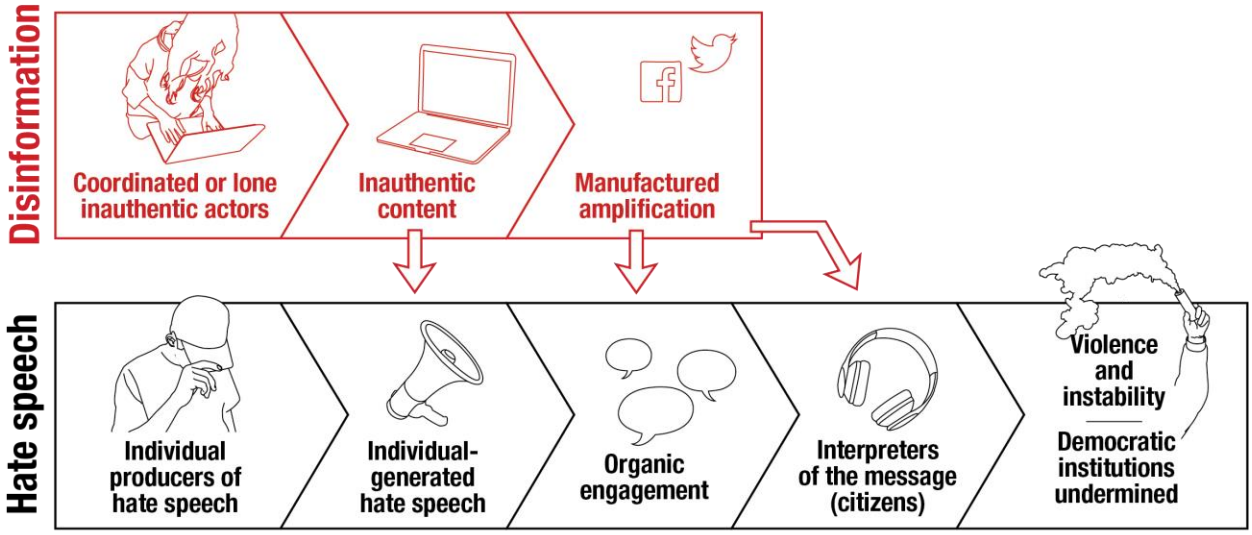
Effective programming to counter or mitigate the effects of hate speech that spreads according to the graphic above can target one or more of the individual components (*actor, message*, *mode of dissemination*, *interpreter* or *risk*)[9] to disrupt the trajectory. For example, in a scenario where the organic, viral spread of hate speech through traditional and social media is heightening the risk of electoral violence, one programming approach, among many possibilities, might include:

TARGETING THE ACTOR	TARGETING THE MESSAGE	TARGETING THE MODE OF DISINFORMATION	TARGETING THE INTERPRETER	MITIGATING RISK
Review of the legal framework for identifying and sanctioning producers of hate speech. Development of more effective investigation and enforcement tactics to prosecute offenders within the bounds of the law.	Empower public actors, such as election management bodies (EMBs), to model inclusive practices and speak out against hate speech. Provide alternative narratives by elevating the voices of marginalized groups.	Codes of conduct for media actors to prevent the spread of hate speech through traditional media sources. Pressure on social media companies to slow the spread of viral hate speech, or enforce their hate speech community standards and remove hateful content.	Civic education that promotes tolerance. Awareness campaigns that educate the public about hate speech.	Coordinated security planning by multiple agencies to mitigate the risk of electoral violence.

The next graphic builds on Figure 1, depicting how the trajectory of hate speech (earlier graphic duplicated in black) is altered and intensified when a disinformation campaign is deployed. As the new graphic illustrates, inauthentic actors deploying disinformation tactics inject fabricated and manipulated content into the pool of hate speech already in circulation, altering the *message* and *modes of dissemination.* The effect of a disinformation campaign is depicted as a second layer (in red) in the chain of harm, and the text below the graphic depicts the process by which these disinformation efforts amplify the risk and magnitude of violence and democratic erosion. Definitions of the individual elements included in this image are available in the glossary at the end of this brief.

[9] This model adapts ideas from Wardle (2017), which conceptualizes the *elements* of information disorder as agent, message and interpreter and the *phases* as creation, production and distribution.

Figure 2: Disinformation-Amplified Hate Speech



ACTOR	MESSAGE	MODE OF DISSEMINATION	INTERPRETER	RISK
The perpetrators of disinformation campaigns may be lone individuals motivated by ideology, but increasingly are coordinated foreign or domestic actors intending to suppress political participation, create confusion and distrust, and polarize the electorate to spread content virally - all in the service of undermining societal cohesion and democracy, promoting the political rise or preservation of an individual or group, or financial gain.	The creation of content intended to deceive the public (e.g. junk news, deepfakes) amplifies and reinforces narratives already in circulation. These messages are created to receive maximum visibility and calculated to play on the cognitive biases of those who engage with them. Artificial intelligence (AI) enables the creation of increasingly convincing manipulations of images and content.	Through paid engagement and networks of coordinated social media accounts (human, bot and hybrid), inauthentic content is unleashed to flood the information space. As this content gains the appearance of credibility through high levels of (artificial) engagement, users become increasingly likely to re-share content and messages may jump to traditional media and spread through word of mouth. Additionally, algorithms and AI take advantage of vast troves of personal data to enable the targeted dissemination of messages in ways that maximize their persuasiveness to particular audiences.	Manufactured amplification can make hate speech messages seem more widely held and prevalent than they are, emboldening normally passive citizens or shifting the electorate's perception of popular opinion and degree of hostility toward the targets of hate speech. Citizens become more likely to perceive threats to the integrity of political and electoral processes.	Risks multiply as citizens' ability to distinguish true and false narratives diminishes and a sense of urgency, unfairness and threat rises.

Illustrative examples of disinformation campaigns using hate speech as a tactic include:

- Political parties financing troll farms to distribute fake sexualized content that harasses, discredits and humiliates female candidates in the opposition.
- A coordinated campaign organized by military elites that utilizes false social media accounts, troll farms and other disinformation tactics to amplify existing societal divisions and incite violence through the demonization of an ethnic minority group.
- A voter suppression effort targeting a racial or ethnic minority that applies a coordinated disinformation campaign to sow distrust in democratic institutions and political processes.
- Politically extreme organizations successfully mobilizing voters to turn out against a gubernatorial candidate by manufacturing and artificially amplifying offenses against the majority religion.

Visualizing the relationship between hate speech and disinformation as we have done above makes apparent the increased number of intervention points available for anti-disinformation programming – while also illustrating why some programming techniques effective at countering ideologically motivated hate speech might be less impactful when technology-fueled disinformation tactics have altered the information landscape.

The ultimate goal of such programming is to stop the perpetrators of hate speech and disinformation (the *actors*) from causing violence or from undermining faith in democratic processes and institutions (the *risk*). To that end, interventions can disrupt the chain of harm at any point (*actor*, *message*, *mode of dissemination*, *interpreter* or *risk*). For example, if programming could create extremely savvy interpreters who are entirely impervious to disinformation and hate speech, no other intervention would be needed along the chain of harm. This is patently unrealistic, of course, so effective approaches will need to attack at multiple intervention points. Though some intervention points will be more or less impactful, the problem does not have to be equally neutralized at every phase along the chain to mitigate the ultimate risk. Importantly, when a disinformation campaign is present in addition to hate speech, programming must take into account both layers of the chain of harm.

As disinformation-amplified hate speech becomes the norm in political discourse in many countries, approaches to countering hate speech should also address the new media context. If hate content is generated by inauthentic actors motivated purely by political or financial calculations, making individuals aware of what constitutes hate speech, why it is wrong and how they could be punished for engaging in it might not address the full scope of the problem. Programming interventions to address hate speech can be fruitfully combined with interventions that also target the disinformation aspect of the problem. Below is a chart of some illustrative programming examples (there are many others that would be appropriate in various contexts), intended to highlight how interventions that target hate speech alone might differ from interventions that target disinformation campaigns utilizing hate speech as a tactic:

	TARGETING THE ACTOR	TARGETING THE MESSAGE	TARGETING THE MODE OF DISINFORMATION	TARGETING THE INTERPRETER	MITIGATING RISK
Hate Speech	Review of the legal framework for identifying and sanctioning producers of hate speech. Development of more effective investigation and enforcement tactics to prosecute offenders within the bounds of the law.	Empower public actors, such as EMBs, to model inclusive practices and speak out against hate speech. Provide alternative narratives by elevating the voices of marginalized groups.	Codes of conduct for media actors to prevent the spread of hate speech through traditional media sources. Pressure on social media companies to slow the spread of viral hate speech, or enforce their hate speech community standards and remove hateful content.	Civic education that promotes tolerance. Awareness campaigns that educate the public about hate speech.	Coordinated security planning by multiple agencies to mitigate the risk of electoral violence.
Disinformation	Political finance regulation to prohibit domestic political actors from engaging in the production and dissemination of inauthentic content.	Advocacy to demand greater transparency from social media companies on inauthentic content circulating on their platforms. Research and monitoring of evolving trends in inauthentic content production.	Identification and removal of inauthetic coordinated behavior on social media platforms.	Civic education that promotes media literacy. Public service announcements that educate the public about disinformation campaigns.	Crisis planning for EMBs on how to address and mitigate the impacts of disinformation during critical parts of the electoral cycle.

Concluding Thoughts

The modern disinformation problem presents a clear threat to the information ecosystems that underpin the health of democratic institutions and processes. The calculated amplification of hate speech is a particularly virulent tactic used by some disinformation actors to promote agendas that are antithetical to democratic values. As this brief has outlined, evolving technologies – and an understanding of human cognitive biases that are ripe for algorithmic exploitation – offer these actors the ability to create sophisticated and powerful disinformation campaigns that increase and intensify hate speech already circulating in a political or electoral context.

While hate speech and disinformation are intimately related concepts, understanding the nuance and interplay between them is essential to designing responsive programming. IFES programming in this arena builds on longstanding relationships with a range of actors around the globe who have a role to play in countering hate speech and the spread of false narratives. This programming is continually adapting to meet the challenges of a rapidly evolving new media and technology environment in order to equip election administrators and electoral stakeholders to intervene at multiple points of the disinformation chain of harm.

Glossary: Defining Elements of a Disinformation Campaign

We have included this brief glossary as an additional reference to guide this complex and evolving discussion over the integrity of political information and discourse. These definitions, adapted from an array of leading sources in the understanding of contemporary disinformation, do not comprise an exhaustive list, however, and the tactics used in disinformation campaigns will continue to evolve.

Coordinated or Lone Inauthentic Actors

Influence campaigns are "actions taken by governments or organized non-state actors to distort domestic or foreign political sentiment, most frequently to achieve a strategic and/or geopolitical outcome."[10] Influence campaigns increasingly deploy an array of disinformation tactics with the goal of manipulating public opinion and undermining the integrity of the information environment.

Coordinated inauthentic behavior is when groups or individuals work together to mislead others about who they are or what they do. The identification of this behavior is not dependent on the content that is shared by these actors, but rather by the deceptive behaviors that they use.[11]

Internet trolls are human users on internet platforms who intentionally harass, provoke, or intimidate others, often to distract and sow discord. Trolls can act as individuals, and in this capacity share many characteristics with individual perpetrators of hate speech. However, trolls can also engage in coordinated inauthentic behavior.

Inauthentic Content

Junk news includes the publication of propaganda and ideologically extreme, hyperpartisan or conspiratorial political news and information under the guise of providing credible information. The term includes news publications that present verifiably false content or commentary as factual news.[12]

Deepfakes are digitally altered images and videos that use artificial intelligence to combine real source material with manufactured content to create hyper-realistic portrayals of individuals saying or doing things that did not occur.

Manufactured Amplification

Computational propaganda is "the use of algorithms, automation, and human curation to purposefully distribute misleading information over social media networks. Computational

[10] Wardle (2017), page 16.
[11] "Coordinated Inauthentic Behavior Explained," Facebook Newsroom, Dec 6, 2018, https://newsroom.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/.
[12] This definition is adapted from the Oxford Internet Institute, https://newsaggregator.oii.ox.ac.uk/methodology.php.

propaganda involves learning from and mimicking real people so as to manipulate public opinion across a diverse range of platforms and device networks."[13]

Bots are simple computer codes that can simulate human beings and make posts online. Botnets are the coordinated efforts of multiple bots.

Content or click farms are commercial enterprises that employ individuals to generate fraudulent profiles, posts and "likes" in order to promote specific narratives online. Coordinated efforts to direct the attention of internet trolls toward particular targets or in promotion of certain messages can use the same model as content farms, and are referred to as "troll farms."

[13] Woolley, Samuel C. & Philip N. Howard, "Computational Propaganda Worldwide: Executive Summary," page 6, http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/06/Casestudies-ExecutiveSummary.pdf.

Global Expertise. Local Solutions.
Sustainable Democracy.

IFES | 2011 Crystal Drive | 10th Floor | Arlington, VA 22202 | www.IFES.org