HOW CIVIL SOCIETY CAN COMBAT MISINFORMATION AND HATE SPEECH WITHOUT MAKING IT WORSE

By Dr. Joan Donovan

THE MEDIA MANIPULATION CASEBOOK

HOW CIVIL SOCIETY CAN COMBAT MISINFORMATION AND HATE SPEECH WITHOUT MAKING IT WORSE

The 3Ds of Disinformation: Detect, Document, and Debunk

In the battle to combat misinformation, researchers have offered clear advice for how journalists should cover and debunk it, but we have very provided little guidance for how civil society should counter media manipulation and disinformation campaigns. The lack of attention to civil society responses is a major gap in the research and it's becoming increasingly clear that the guidance for journalists does not translate easily to civil society. At this time, we need all hands on deck to ensure a free and fair election. In this document, I explore a set of potential strategies to be used specifically by civil society organizations (CSOs) to mitigate the harms of misinformation.

The following recommendations are based on the mixed-methods investigative work of the Technology and Social Change (TaSC) research team at Harvard's Shorenstein Center. Before delving into the specifics of the recommendations, it's important to note how the research that informs them is conducted. At TaSC, we use the methods of digital ethnography to detect, document, and debunk media manipulation disinformation campaigns. In the detection phase, we begin by looking closely for the suspected misinformation across different websites and social media platforms. If we suspect a website, post, or video contains misinformation, we then investigate if the account or website is legitimate and representing itself accurately. We use various open-source investigation tools available on the web to document signs of imposters, including assessing the history of the accounts in question, the degree of automation on a specific topic, and the networks of actors sharing the suspicious materials.

Once we have documented a pattern of malicious behavior, we create a timeline of events for deeper analysis. During our investigation, we gather evidence by taking screenshots or a screen recording of everything we encounter. If it is misinformation, there is a chance that tech companies may delete content without notifying anyone. The more evidence we have, the better we can make our case.

The more evidence we have, the better we can make our case.

Debunking does not happen during the investigation process. It happens after we piece together the evidence and make inferences about what happened. Following the framework of the *Media Manipulation Life Cycle*, we draw together collected materials to analyze timelines, behavioral patterns, and the broader context of the disinformation, including how it attaches to hot-button wedge issues, which are contested political issues that often have to do with the redistribution of rights, resources, or representation. Only after we have established a timeline and corroborating evidence are we able to assess the impact of misinformation on specific issues and communities and strategically debunk disinformation, either in concert with journalists and civil society, or directly through our research reports.

While this long process makes sense for research, like journalists, CSOs often do not have the luxury of time. Confronting misinformation while working with partial knowledge is difficult, but not impossible. CSOs and activists will often know more about the local and specific context where misinformation is having a negative impact. In some situations, organizers may see a disinformation trend in small online communities, or they may be the target of attacks. Frequently we see communities of color impacted by racialized disinformation campaigns, which employ the strategic use of falsified racial or ethnic identities and/or focus on race as a wedge issue. These tactics were used against communities of color in 2016 to suppress their votes and electoral power. Responding to disinformation can be complicated, because it can amplify misinformation and further the agenda of media manipulators.
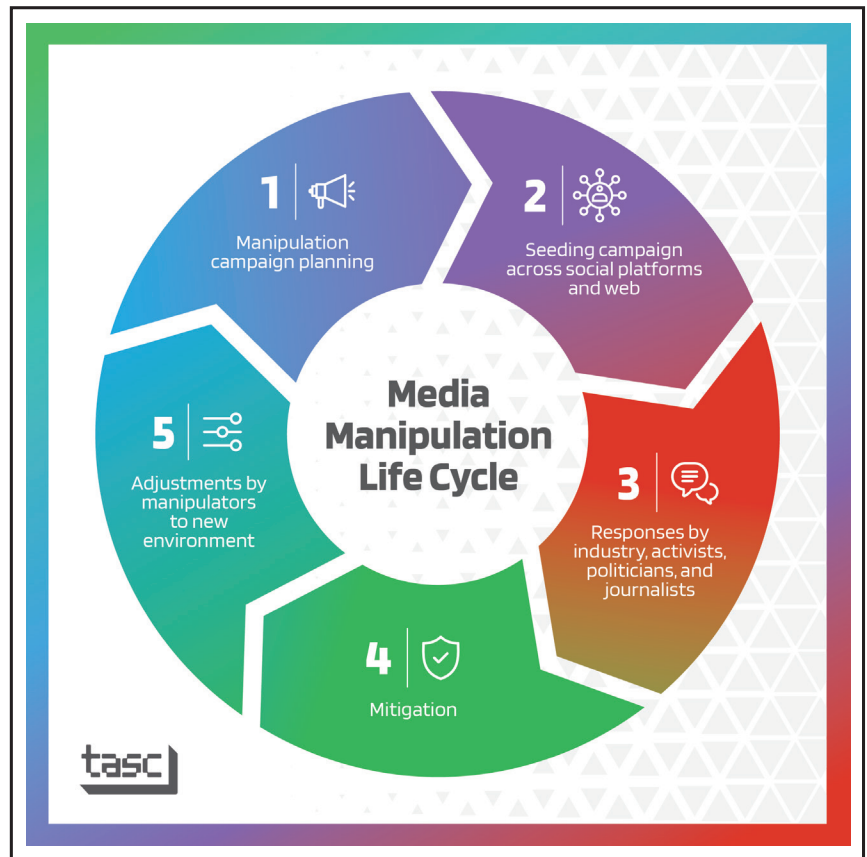
Therefore, we are developing research on several case studies where CSOs and/ or small groups employed different strategies to counter misinformation. It is often difficult to get important information to rise above the cacophony of misinformation.

Civil Society Response: Six Strategies for Countering Misinformation and Hate Speech

1   Connected Communities

Before taking up any kind of strategic response to disinformation, social networks should be thought of as connected communities and not as audiences. Telling people what to think, as if they were merely receptacles or passive audiences, leads to disengagement. In some instances, organizing a response to disinformation can take on the look and feel of a campaign, but this is not an effective approach to disinformation, as it undercuts the possibility for communities to collectively learn, counter, and adapt to disinformation targeted at the community.

The rationale for any response to disinformation should be grounded in facts about the impact and harms of particular trends in disinformation.

Begin by assessing what types of disinformation are present and determine what kind of actions to take. Reactions could be as simple as replying to an online post to say, "This conversation doesn't belong in this group" or to link to a reputable source that disputes the disinformation. Flagging posts is an underutilized tool available on most platforms.

Brandi Collins-Dexter, a Senior Fellow at Harvard Kennedy School's Shorenstein Center, researches how Black online communities create and maintain norms of engagement across social networks. For each online community, participation and trust is negotiated in terms of safety and protection from harassers, trolls, and even well-meaning journalists. Maintenance of internal norms of discussion is an important factor when addressing misinformation. In places where there are no community moderation rules, misinformation thrives. When dealing with communities that may be spreading misinformation out of concern and fear, approaching individuals with an ethic of care can quickly dissipate growing tensions. If that does not work, moderators should consider removing posts or accounts.

When dealing with communities that may be spreading misinformation out of concern and fear, approaching individuals with an ethic of care can quickly dissipate growing tensions.

When dealing with online communities designed to spread misinformation, other strategies are needed. In 2020, Reddit deleted a number of subreddits that were used to plan coordinated harassment campaigns and were a safe haven for hate speech. This did not happen solely because the company assessed the risk and made a decision, but because groups like Change the Terms and many others applied public pressure by noting that misinformation hives, like Reddit's the_donald, do damage to other communities. As content moderation has become a crucial aspect of maintaining an online community, the safety,

integrity, and protection of well-functioning connected communities is a paramount concern for CSOs.

2 Fact/Fallacy/Fact or "The Truth Sandwich"

Rhetorically, how you say something is just as important as what you say.

Often disinformation is delivered memetically, i.e. in the form of short, memorable, and pithy slogans. An untrue statement, such as, "Vaccines cause autism," is particularly difficult to disrupt once it's been repeated. Refuting a statement like this requires breaking up the psychological impulse to remember something so sticky and then replacing it with what is true. A rebuttal should take the form of a truth sandwich, such as: "Vaccines don't cause autism. While the myth persists that vaccines cause this condition, doctors across the globe have proven that vaccines do not cause autism and are a benefit the whole of society."

Another way of constructing the truth sandwich can include highlighting the political agenda of the group pushing lies. In that way, a message dispelling myths about vaccines could read: "Vaccines don't cause autism. This myth is perpetuated by anti-vaccine activists and does not line up with scientific facts about public health. Doctors across the globe have proven that vaccines do not cause autism and are a benefit the whole society."

The truth sandwich can be used in any number of situations and is best employed to counter broad thematic trends in misinformation.

3 Prebunking

In some cases, disinformation is predictable. In order for a lie to take hold in public discussion, audiences usually have to encounter it numerous times. Prebunking is an offensive strategy that refers to anticipating what disinformation is likely to be repeated by politicians, pundits, and provocateurs during key events and having already prepared a response based on past fact-checks.

Across every wedge issue, there are different kinds of disinformation that continue to pollute our information ecosystem. For example, the upcoming presidential debate will be rife with misinformation, as Trump and Biden do not align on any issue. As a result, the truth will be contorted to fit partisan talking points, especially as it relates to pandemic response, immigration, religion, racism, and white supremacy. Knowing the most common misperceptions of each candidate and potential rebuttals with links to truthful and authoritative information can stop disinformation from spreading, especially if the prebunk is published closely following the misinformation.

For example, Daniel Acosta Ramos at First Draft News was a fact-checker during the controversial 2018 election in Venezuela, where Nicolás Maduro claimed victory and is now considered a contested ruler. In anticipation of election disinformation, Daniel, along with other fact-checkers, prepared a list of myths and prebunks, so journalists could quickly counter any misinformation about voting or candidate's platforms. As the situation with the election evolved, fact-checkers were ready with easy-to-share rebuttals for common disinformation trends.

Prebunking efforts require a keen eye for details. Because successful disinformation campaigns tend to have a kernel of truth that anchors the narrative, **disinformation parasitically attaches itself to the details of a particular issue**. Disinformation does the most damage to public discourse when it has leeched on to a complex topic, which a nuanced rebuttal is needed. Prebunking is an effort to get ahead of misinformation and establish credibility and trust before the disinformation narrative can take hold. Being aware that certain themes are likely to come up provides an added advantage to communities who may otherwise get bogged down by debating disinformation.

> **Because successful disinformation campaigns tend to have a kernel of truth that anchors the narrative, disinformation parasitically attaches itself to the details of a particular issue.**

Doing prebunks is traditionally in the domain of journalists, but truth needs as many advocates as possible right now. In order to effectively prebunk misinformation, CSOs should follow this 5-step process:

1. Take a look at fact checking websites and databases to get a sense of the trends in misinformation.
2. Map out which misinformation trends are popular in politicians' stump speeches.
3. Find additional source material with the facts about the misinformation likely to be repeated and create some content that dispels these themes.
4. Prepare your social networks for the high potential for misinformation on particular themes by sharing select prebunks. *Repeating misinformation can have unintended negative effects. It is best to utilize the truth sandwich model, cite reliable evidence, link to fact-checks, and state the truth upfront.
5. Prebunks can also be reused as debunks. During an event like a debate, election night, or other major breaking news, when false claims are made, posting the correct information alongside the misinformation quickly can make a difference. Speed matters as misinformation narratives can take hold quickly on social media when little factual information is available.

CSOs can issue prebunks as press releases or social media content ahead of events, but this might get tricky if a certain topic does not come up. Be cautious because prebunks that do not come to fruition could inadvertently seed misinformation.

4 Distributed Debunking

In many situations where disinformation starts to trend on social media, it is because those who are pushing the disinformation are battling it out with those who are debunking it. Engaging with disinformation often makes it gain more traction within search and trending algorithms, because these technologies cannot tell the difference between truth and lies. It is, therefore, usually not worth the time and resources to counter disinformation, because those efforts serve to amplify the lies. However, in situations where disinformation has scaled from the depths of message boards and blogs to mainstream media and triggered responses from law enforcement, politicians, and other newsworthy groups, it has then reached a point where a response may be needed.

For civil society, response should be separated from reaction. Reactions, like retweeting a journalists' debunk, can sometimes be enough to show awareness and correct the issue. Responses, though, take more time and should be thought about strategically. Often, we see lies go viral because people are participating in a collective process of distributed debunking, where it is more important to circulate the truth than to ignore the disinformation. In most cases, activists and CSOs know more about the context and impact of disinformation, and their responses are crucial for documenting the truth.

For example, #AntifaFires began trending on Twitter as wildfires displaced many people in Oregon and Northern California. The disinformation is possibly attributed to a misunderstanding on a police scanner that was reporting on controlled burns by "BLM," Bureau of Land Management, where BLM was confused with Black Lives Matter. Then a series of right-wing blogs and social media accounts began to circulate rumors that Antifa and Black Lives Matter protesters were arsonists. A significant source for this disinformation came from the right-wing publication Law Enforcement Today, which published a story with the headline, "Sources: Series of wildfires on the West Coast may be 'coordinated and planned' attack," which gained nearly 400,000 interactions on Facebook. The publication has since changed the title to "Arson arrests made across the west coast as fires rage on," in reaction to fact-checks debunking their unsubstantiated claims. Nevertheless, disinformation proliferated across all social media platforms, leading local police, sheriffs, and even the FBI to issue statements saying that Antifa did not set the fires. All the while, activists, organizers, and civil society groups stepped up to further debunk these false claims.

Because the damage had already been done to seed the disinformation into the media ecosystem, distributed debunking as a strategic response became a way to show solidarity, while also providing a moral counterweight to disinformation. One important caveat, though: simply

In most cases, activists and CSOs know more about the context and impact of disinformation, and their responses are crucial for documenting the truth.

posting that you don't think this is true will only add to the cacophony of noise that boosts the disinformation. Distributed debunking needs to be organized to be effective. In order to become a powerful tool to collectively shut down lies, debunks should include a link to a reputable source and follow the models of the "truth sandwich" or "humor over rumor" (described below in Section 6).

5 Localize the Context

All disinformation is local. CSOs are often on the ground in communities and therefore in the best possible position to provide this context. When directly debunking information, bear the local community in mind, and if possible, share your local knowledge with journalists and other CSOs for whom the information may be crucial.

Knowing how a disinformation campaign impacts and causes harm in communities, both online and off, is critical information for journalists covering particular beats. Keeping track and monitoring disinformation is hard work, but diligent and sustained documentation of disinformation does yield patterns, from which successful counter strategies can be developed. Media manipulators and disinformation agents are not as numerous as once thought. However, manipulators are having an outsized impact on the news media because they employ a diverse set of tactics and do whatever it takes to keep their campaign alive.

For example, if we look at the patterns of disinformation related to the overblown claims that Antifa are flying from town to town wreaking havoc and weaponizing cans of soup and tuna, then the allegation that Antifa were setting wildfires makes perfect sense as a disinformation campaign. In the days before this rumor kicked off, the US was rapt by images of the orange skies over the West Coast. Stories about the wildfires were rapidly gaining popularity, where attention to the causes and consequences of climate change were taking up oxygen in an already suffocating media ecosystem. In order to regain control of the news media narrative, falsely associating the wildfires with Antifa and Black Lives Matter was a political opportunity to shift attention, parry blame, and then step back and watch the chaos unfold. If the breaking news was not about the wildfires, perhaps a rumor like this would never have taken hold.

Crucially, over the summer, it was protesters in Portland who first called attention to unmarked Federal agents arresting people at protests. While protesters posted several suspicious videos to social media, they could not prove what they knew was happening. This local knowledge from protesters was then amplified by CSOs through their social media networks to keep the attention on the issue. Reporters then picked up the story, filed FOIA requests, and sought corroborating evidence

from local businesses and others. All the while, government agents dodged questions about their operations in Portland, which allowed misinformation to proliferate. The combined efforts of journalists, CSOs, and protesters to force accountability and transparency are integral for calling out those in power for their actions.

6 Humor Over Rumor

The psychology of misinformation is complex, but tends to begin by triggering emotional reactions and confirmation bias. Misinformation thrives in environments that are charged up by outrage, fear, and anger. It is especially salient when the group sharing the misinformation hold similar political and cultural points of view, largely because they will offer few challenges or criticisms of the misinformation. Additionally, people are inclined to share information if they feel that it is somehow being suppressed or is not going to be covered by the media. When these set-up conditions are reflected in online communities, especially Facebook groups or message boards, misinformation can be difficult to fight. Like a rising tide headed to the shore, some damage is inevitable, but there are actions that can be taken to guard against this.

Humor over rumor is a community strategy employed by the digital ministry of Taiwan to respond to misinformation quickly and memorably. For example, at the beginning of the pandemic, a rumor began to spread on social media that toilet paper was made of the same material as masks. Anticipating a run on toilet paper, the digital ministry kicked into gear and created some funny memes dispelling the rumor and showing that the original misinformation came from the producers of toilet paper.

"We only have one pair of buttocks" meme from Taiwanese Premier Su Tseng-chang.

Audrey Tang, the Digital Minister of Taiwan, and I discussed the role memes play countering heightened emotional reactions to misinformation. In particular, the pandemic has produced a toxic mix of outrage and isolation that has driven some to ponder conspiracy theories. The use of memes to check coronavirus misinformation in Taiwan has been successful for two reasons: humor and speed. Using a dog named Zongchai as a spokesperson, the digital ministry employs comedians and other artists to quickly respond to emerging rumors. The speediness of the response is crucial to overtaking the misinformation or rumor before it becomes too widespread. When done right, the humorous rebuttal attaches itself

to the misinformation, and is found everywhere the rumor is spreading. By making these fact-checks funny, it encourages organic distribution. This is crucial, as meme distribution cannot be top-down; ordinary citizens spearhead the meme war.

Experiments like this have yet to be tried in the US. In order to quell dangerous misinformation about the pandemic, humor can be a strong mechanism for delivering the truth. In the run up to the 2020 election, humor and irony can also be used to help get out the vote, particularly as misinformation attacking election integrity seeks to bring about voter apathy.

The "spokesdog" for Taiwan's Ministry of Health & Welfare.

Conclusion

[Now is the time for experiments and action](). Rather than wait for tech companies to mitigate the damage caused by their products, civil society must try new strategies for fighting media manipulation and disinformation. Our current information ecosystem was not designed to serve the interests of the broader public, nor was it designed to protect communities caught in the crosshairs of a disinformation campaign.

Truth needs an advocate.

About the Author

Joan Donovan, PhD, is the Research Director of Harvard Kennedy's Shorenstein Center and the Director of the Technology and Social Change (TaSC) Research project. Dr. Donovan leads the field in examining media manipulation and disinformation campaigns. Her research specializes in Critical Internet Studies, Science and Technology Studies, and the Sociology of Social Movements.

THE
CASEBOOK