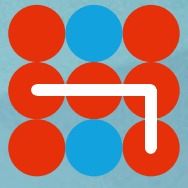
COI HYBRID INFLUENCE

Hybrid CoE Working Paper 11

JULY 2021

Disinformation 2.0: Trends for 2021 and beyond

ISABELLA GARCIA-CAMARGO AND
SAMANTHA BRADSHAW



Hybrid CoE

Hybrid CoE Working Paper 11

Disinformation 2.0: Trends for 2021 and beyond

ISABELLA GARCIA-CAMARGO AND
SAMANTHA BRADSHAW

Hybrid CoE Working Papers cover work in progress: they develop and share ideas on Hybrid CoE's ongoing research/ workstrand themes or analyze actors, events or concepts that are relevant from the point of view of hybrid threats. They cover a wide range of topics related to the constantly evolving security environment.

COI Hybrid Influence looks at how state and non-state actors conduct influence activities targeted at Participating States and institutions, as part of a hybrid campaign, and how hostile state actors use their influence tools in ways that attempt to sow instability, or curtail the sovereignty of other nations and the independence of institutions. The focus is on the behaviours, activities, and tools that a hostile actor can use. The goal is to equip practitioners with the tools they need to respond to and deter hybrid threats. COI HI is led by the UK.

The European Centre of Excellence for Countering Hybrid Threats tel. +358 400 253800 www.hybridcoe.fi

ISBN (web) 978-952-7282-88-5
ISBN (print) 978-952-7282-89-2
ISSN 2670-160X

July 2021

Hybrid CoE is an international hub for practitioners and experts, building Participating States' and institutions' capabilities and enhancing EU-NATO cooperation in countering hybrid threats, located in Helsinki, Finland.

Contents

Executive summary

Russian interference in the 2016 US presidential election drew attention to the many ways that social media can be leveraged for widescale information operations by nation-state actors. Since then, however, these campaigns have evolved: Rather than a large foreign effort to undermine the integrity of the election, the new digital threats to the 2020 US presidential election came from domestic users who co-opted social media platforms to spread disinformation about a "stolen election". Groups like #StopTheSteal – coordinated by many authentic American users – seeded false claims about illegal voting on Facebook and Twitter.[1] Conservative and far-right livestreams on YouTube amplified conspiracies about rigged voting machines to audiences that rival television or broadcast shows.[2] Users on TikTok and Instagram posted stories about missing, destroyed, or uncounted ballots on election day – many of which would disappear after 24 hours.[3] And on fringe or alternative platforms like Parler, Gab, and 4chan, users left comments inciting the "rape, torture, and the assassination of named public officials and private citizens", rhetoric that ultimately led to an insurrection at the US Capitol on 6 January 2021.

In this Hybrid CoE Working Paper, we evaluate how disinformation has evolved from 2016 to 2020. We define "disinformation" as the purposeful spread of false, misleading, or exaggerated content online, or the use of fake accounts or pages designed to purposefully manipulate or mislead users. Using the 2020 election in the United States as our case study, we examine how the actors, strategies, and tactics for spreading disinformation have evolved over the past four years, and discuss the direction that future policymaking should take to address contemporary trends in information operations. Our data and examples are drawn from the research and analysis we conducted as part of the Election Integrity Partnership, facilitated by the Center for an Informed Public, the Digital Forensic Research lab, Graphika and the Stanford Internet Observatory. Here, we identified ten key trends about 2020 disinformation campaigns:

1) Disinformation actors made use of **hard-to-verify content,** which is content that is not falsifiable through fact-checking. Hard-to-verify content can evade takedowns because it blurs the line between truth and fiction, with users framing disinformation as stories they heard from friends or asking questions about known conspiracies or disinformation.
2) Rather than bots or fake accounts amplifying disinformation, most viral disinformation was **top-down** and was amplified by influencer accounts with millions of followers. As platforms consider speech from influencers – like politicians – newsworthy, content produced and shared by influencer accounts is considered a public good and has not been moderated as strictly as that from other non-newsworthy accounts.
3) Disinformation narratives were not spread on a single platform but often appeared **across multiple platforms.** This made it difficult to remove conspiracy theories or disinformation as they emerged online, since information operations often extended beyond the remit of a single platform to respond.
4) Social media platforms did not take action quickly enough against **groups and online communities** that encouraged and coordinated the authentic bottom-up spread of disinformation. These entities were the primary drivers of key disinformation campaigns, which spilled into offline action, including #StopTheSteal.
5) While recommendation algorithms are foundational to the business models of social media platforms, **algorithms can also amplify**

1 Silverman, Mac, and Lytvynenko, 'Facebook Knows It Was Used To Help Incite The Capitol Insurrection'.
2 Bradshaw et al., 'Election Delegitimization'.
3 Paul, 'TikTok: False Posts about US Election Reach Hundreds of Thousands'.

disinformation and reinforce viewpoints within certain groups of users. Some election delegitimization narratives were picked up and recommended by platform algorithms.

6) Many disinformation actors leveraged **images, videos and ephemeral content** to create highly engaging posts, and elicit strong emotional responses in stories that often disappeared after 24 hours. Researchers, fact-checkers, and even social media platforms themselves found it difficult to moderate or track the origin, spread and impact of this specific disinformation tactic.
7) Video streaming platforms which were used to **livestream** content blended disinformation and conspiracy into online talk shows, often broadcasting to millions of people in real time. The real-time nature of this content combined with the fact that shows could be an hour-long segment with only a brief mention of a conspiracy theory introduced a range of moderation challenges that platforms were not prepared to address.
8) Narratives moderated by mainstream platforms would reappear in **extremist communities** online, which drew a large number of followers who were similarly removed from mainstream platforms. Extremist communities, like Parler, Gab, 4chan or 8chan, take a laissez-faire approach to content moderation, which allowed for disinformation and conspiracy to continue to find a home online.
9) Certain communities of users – like diaspora communities – communicated through **encrypted platforms,** where harmful disinformation about the legitimacy of the election spread. The encrypted nature of the platforms made this content much harder to track. Additionally, the organic trust engendered within these closed groups creates a difficult environment for an external entity to fact-check any misleading election-related claims that may be discussed.
10) Although there were concerns about the use of AI-generated technologies to generate deepfakes – or completely fictional videos that look realistic – **the use of AI remained largely untapped** during the 2020 US presidential election. Sometimes, AI would be used for deception – to create fake profile photos of a person who does not exist to make inauthentic accounts appear real. However, no large-scale disinformation campaigns or viral content relied primarily on the use of these technologies.

Given growing concerns over the misuse of social media platforms, the companies themselves have adopted a series of measures designed to limit the harms of disinformation. These strategies have ranged from improving both human and automated content moderation capabilities to investing in fact-checking initiatives, or labelling information sources to help users distinguish trustworthy entities from dubious ones. [4] In the lead-up to the 2020 US presidential election, social media platforms also introduced election-specific policies that addressed content meant to stagnate the spread of election-related mis- and disinformation, such as claims delegitimizing election results or calls to interfere with voting operations through violence.[5] However, despite these initiatives, the spread of disinformation continues to disrupt democratic deliberation online. We identify four areas in which policies were inadequate to address the realized dynamics:

1) The most salient disinformation narratives were **coordinated by authentic users**, rather than fake, inauthentic accounts. While platforms have developed and implemented policies designed to target *inauthentic* behaviour, decisions were much slower and more difficult when real users were involved.
2) While platforms have policies targeting election disinformation, some disinformation **evaded platform policies and takedowns.** In particular, some viral disinformation was not clearly falsifiable and did not always appear harmful or use the vocabulary of violence. Since 2016, platform moderation has become much better at identifying and removing "fake news" and labelling information that has been

4 Taylor, Walsh, and Bradshaw, 'Industry Responses to the Malicious Use of Social Media'.
5 Election Integrity Partnership, 'Evaluating Platform Election-Related Speech Policies'.

fact-checked by third parties as false. However, disinformation narratives have adopted new strategies to present falsifiable information as stories from friends, making it hard – if not impossible – to verify the veracity of the statement. Other users would "just ask questions" about conspiracies or frame disinformation narratives as an open question. These dynamics made it difficult for platforms to quickly demote and remove disinformation online.

3) **Disinformation was networked** across many platforms, including unmoderated or encrypted spaces, making it hard to track and remove harmful content at scale across the information landscape. Narratives did not emerge on one platform and stay there; rather they would be amplified and reinforced across multiple mainstream platforms. Platforms did not have standardized best practices or coordination across them to deal with disinformation as it emerged in various forms across different information ecosystems.
4) Certain narratives about election delegitimization continued to be algorithmically amplified by platforms whose business models continue to **incentivize the viral spread of disinformation**. While many of these problems were brought to public attention after the 2016 US presidential election, researchers still do not have access to the data to measure the impact of algorithmically amplified disinformation, particularly when algorithms promote certain kinds of stories to users based on data about their political values or beliefs.
5) Although the use of AI and other innovations in technology were limited during the 2020 US presidential election, **AI challenges on the horizon** continue to loom over platforms and their ability to detect inauthentic content and behaviour. Platforms have developed policies to label AI-generated content, and continue to develop technologies to identify content and images that are generated with AI tools. However, these technologies and policies have not yet been put to the test, as many of these AI tools were not widely leveraged by disinformation actors.

The evolving threat landscape: New trends in digital disinformation

Social media platforms are used by a variety of actors to spread disinformation about politics. While public attention has largely focused on the role of foreign state actors – like Russia, China, and Iran – in using social media as an extension of geopolitical power, the kinds of actors involved in digital disinformation and coordinated influence operations are diverse. From conspiracy theorists to far-right extremists or political consultancy firms, social media platforms and technologies are being co-opted to spread disinformation in order to undermine the legitimacy of elections and democracy more broadly.

While increased public attention has drawn a growing number of actors involved in information operations, it has also pressured platforms to take action against accounts and content that breach their terms of service and community guidelines. However, the ubiquity of social media platforms combined with the low cost associated with information operations continues to make it a prime medium for manipulation. Adversaries have learned to evade detection, and "game" the technologies and policies in place designed to detect and limit the spread of disinformation online. In this section, we describe nine key trends for 2021 and beyond.

Hard-to-verify content

Many disinformation actors have identified strategies to deliver highly persuasive content that evades platform moderation. By producing content that is hard to verify – often because it is framed as personal experience or lacking a central claim – narratives can achieve their intended goals without violating the platform rules against misinformation. Among the more successful framing strategies observed during the 2020 election were 'friend of a friend' and 'just asking question' narrative framings.

Friend of a friend narratives

A "friend of a friend" story has long been the basis of rumours, even before the rise of online communications. However, in an online context where information spreads quickly and 'friend' is a loose term, this information-sharing pattern has taken on new vigour. An online narrative which begins with, for example, "My friend just posted this…" or "A friend just told me about…", leverages the implied trust in the referenced offline relationship. To users on the receiving end, stories recounting emotive or real-world experiences may be perceived as more credible or relevant coming from an 'in-network' individual, rather than from an unrelated celebrity, politician or stranger.[6] Even though the specific social context is long collapsed as the narrative is posted and reposted, such that 'My friend' may no longer be one but perhaps six or seven network hops away, a receiving individual has no way to gauge the loss of context and likely will receive the message as a unique personal story that is difficult, if not impossible, to fact- check.

These personal stories are easily spread, copied, and pasted across online platforms by well-intended users, each time bringing a perception of unique origins from a trusted source. This format has led to the viral spread of coronavirus misinformation,[7] as well as the rapid propagation of several false election-related narratives.[8] In both the election and coronavirus misinformation context, the actual friend-of-a-friend posts encouraged individuals to copy-and-paste repost the exact message onto their accounts, in order to 'pass along' the potentially relevant information to their own

6 Guadagno, 'Compliance'.
7 Robinson and Spring, 'Coronavirus: How Bad Information Goes Viral'.
8 Koltai, Nasseta, and Starbird, '"Friend of a Friend" Stories as a Vehicle for Misinformation'.

networks. This 'self-spreading' characteristic of 'friend of a friend' narratives makes them an all the more prominent feature of modern disinformation tactics. Additionally, because of the textual nature of the "friend of a friend" post, this type of misinformation narrative often crosses the boundaries of the major social media platforms into additional spread via email, text, and within private group chats (like on WhatsApp), making it even more difficult to track the narrative's spread and provide corrections.

The cross-platform and unverifiable nature of friend-of-a-friend narratives makes this content extremely difficult to moderate for most platforms. Discussions within close networks are the key to collective sensemaking activities. Even beyond the core function of this dynamic for online interactions, the technical capabilities needed to track when a conversation is 'thoughtful discussion' versus potentially misleading or dangerous online gossip is extremely difficult.

The 'Just asking questions' dynamic

Disinformation actors leverage uncertainty to land their messaging among new target audiences. The 'just asking questions' dynamic preys on the key vulnerability of institutional authorities, which must both instil public confidence in the public infrastructure they must protect or deliver, while navigating the difficult territory of acknowledging the shortcomings of that same infrastructure. In the case of vaccine disinformation, for example, the most accurate interpretation of the vaccine's shortcomings will almost certainly be too complex for the public to understand, while simpler interpretations may discredit the authorities in question as the public begins to poke holes in official explanations.

Seasoned disinformation actors are able to exploit this catch-22 that institutional authorities find themselves in, and gain the trust of unwitting audiences by 'simply asking questions', or pointing out the shortcomings or incongruent explanations from less social media savvy authorities. In his Facebook post discussing 'Restrictions for Vaccinated Americans', Fox News host Tucker Carlson's commentary highlights this dynamic when he asks the following seemingly simple questions:

> *"If this stuff works then why can't we LIVE like it works? What are you really telling us here? It's not a trick question. What is the answer?"* (link).

Publicly available information about the US vaccine rollout plan has been opaque and may be inaccessible to users not familiar with Centers for Disease Control and Prevention (CDC) or Food and Drug Administration (FDA) regulations. By emphasizing uncertainty in combination with the endless stream of 'simple questions', influencers spreading disinformation can create an information void for unwitting users who are now receiving manicured answers to questions they did not even consider asking. Through this process, users shift their trust towards partisan influencers, who are conveniently posed with the exact resolution to the uncertainty that they magnified and exaggerated through the 'just asking questions' schematic.

Similar to the 'friend of a friend' dynamic, just asking questions is not inherently bad or malicious: rather, it is a core part of healthy online conversation. The trouble comes when the questioning dynamic is weaponized, shifting an unreasonable burden of truth in a coordinated manner towards central authorities that have both limitations on their own developing knowledge of the situation, and limited bandwidth to address the ongoing barrage of non-substantive allegations posed as legitimate inquiries. When monitoring for misinformation, these dynamics need to be not only tracked but recognized as what they actually are: obstructive claims aiming to further misleading narratives.

Top-down disinformation

'Top-down' disinformation narratives are most often spread by online influencers, who benefit from platform affordances that allow a few individuals to set the agenda for millions of users. Influencers leverage these platform affordances to create an infrastructure for reach and legitimacy around their messaging. Whether their account represents a politician, a celebrity, a media pundit or a different type of digital native, these users drive the attention of huge audiences and accumulate trust within the communities.

Oftentimes, influencer accounts are 'verified' by their respective platforms. This can create a veneer of legitimacy around influencer content, in which social proof drives further sharing and engenders the trust of the audience receiving the message.[9] Network dynamics also favour influencer accounts: simply by their sheer number of followers, influencers are able to drive the conversation for large swathes of users within their network, and even across different platforms.[10] This outsized reach further accumulates on itself, as platform algorithms potentially drive new users to follow verified accounts that seem 'important' within the network that most aligns with the user's expressed interests. During the 2020 election, repeat spreaders of the most pervasive disinformation narratives were overwhelmingly verified, blue-tick accounts belonging to partisan media outlets, social media influencers, and political figures.[11]

Newsworthiness exceptions add to the difficulties of moderating the impact of influencer dynamics. Facebook was the first platform to introduce a newsworthiness clause to their policies in 2016, determining that content was protected from removal if it was determined to be "newsworthy, significant, or important to the public interest". However, determining what is 'newsworthy' quickly becomes a circular process as the celebrity garnered by high-reach individuals across platforms drives their messaging into the public interest time and again.[12] This vulnerability was most clearly highlighted in platform responses to former President Donald Trump's Twitter activity leading up to the 2020 election. On several occasions, content was marked as violative of Twitter's civic integrity policies for clearly delegitimizing the electoral process with baseless claims, but was ultimately preserved on the platform due to its perceived importance to the public interest.

This dynamic is a global phenomenon, and a trend identified by other researchers working on the study of mis- and disinformation: from Australian celebrity endorsements of 5G conspiracy theories to Bollywood movie stars applauding homeopathy fake cures for COVID-19, the emotional connection between fans and their online idols makes internet personalities a disproportionately influential vector in the dissemination of misinformation online.[13]

Groups and online communities

In contrast to the top-down dynamics of platform influencers, groups and other online communities create the foundation for bottom-up or participatory dynamics, allowing like-minded individuals to engage in collective sensemaking. At the same time, groups, community boards, and similar platform infrastructure democratizes the opportunity of large-scale reach, as users who may not have significant political or social capital are able to rally the online communities they are a part of and spread their message widely through this network.

Facebook groups, Reddit boards, or large Telegram channels that lack a clear moderator or leader display the dynamic characteristics of these communal online spaces. Users engage directly with each other to attempt to create in-group consensus on anything from the mundane to highly emotional topics. Once this sensemaking process has accumulated the in-group support for a specific position, what was once a nondescript user's idea has now achieved the potential influence of an external influencer: the individual users who have engaged with and aligned themselves with the position can now go forth and spread this view to their own networks, or amplify the position when they see it echoed by their peers or like-minded influencers.

Coordination may also result from the collective sensemaking process, which goes beyond content amplification. Coordinated harassment methods such as doxing – which involves publishing private or identifying information about an Internet user with malicious intent – can also be coordinated within these in-group dynamics, especially on more private platforms like 4chan that host online

9 Guadagno, 'Compliance'.
10 Abbas Ahmadi and Chan, 'Online Influencers Have Become Powerful Vectors in Promoting False Information and Conspiracy Theories'; Election Integrity Partnership, 'The Long Fuse'.
11 Election Integrity Partnership, 'The Long Fuse'.
12 DiResta and DeButts, '"Newsworthiness", Trump, and the Facebook Oversight Board'.
13 Abbas Ahmadi and Chan, 'Online Influencers Have Become Powerful Vectors'.

communities.[14] This infrastructure is also key to organizing the real-world action that results from online sensemaking: geographically aligned groups in particular can quickly activate offline action from online narrative consensus. While social media platforms can create opportunities for assembly and protest,[15] they can be used to coordinate harm. This was a notable feature of the Stop the Steal group during the US 2020 election: the main Facebook group behind this movement accumulated 360,000 members before it was removed for violating platform rules.[16] This group created multiple events, planning protests in battleground states, encouraging individuals to help pay for "flights and hotels to send people" from out of state to these protests, and even veering into planning for armed conflict.

Ideas spread within online communities also benefit from the in-group trust that these dynamics create. The existence of the group itself provides a certain feeling of belonging or legitimacy for the idea that the users have congregated around: a "New Yorkers against Lockdown" group creates legitimacy for the users who have signalled their adherence to the belief because the user understands that they are not alone in their opinion. Subsequent ideas shared within a group benefit from the mutual trust of the users: instead of fact-checking every new idea, the interpersonal connections created by the users allow individuals to shortcut towards automatically believing content shared by in-group members.[17] Disinformation narratives spread within strong online communities are therefore particularly difficult to counter, as an out-group member will not be able to surpass the trust that the in-group dynamics have created for the collective.

Case study: participatory disinformation – the convergence of top-down & bottom-up disinformation

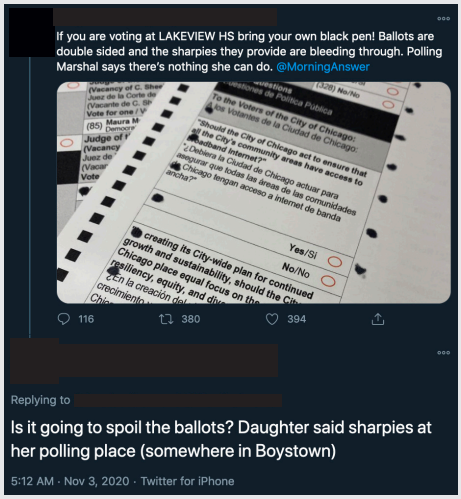
During the 2020 US presidential election, a distinctly ***participatory*** style of disinformation was observed as the mechanism that drove the spread of the most salient online narratives disenfranchising the electoral results.[18] Misleading narratives were not simply fed in a top-down manner from elites to their audiences, but also engaged the 'bottom-up' dynamics of the collective public, in which individual users actively participated in building and amplifying the most pervasive false stories of electoral fraud. Users worked independently to gather 'evidence' for their case through images and videos, developed complicated frames for this evidence, and amplified messages from both their peers and the elites who reinforced their worldview.

Perhaps the most extreme example of this activity came in the form of 'Sharpiegate', which was an online conspiracy suggesting that using a sharpie pen to fill out a ballot invalidated one's vote. Sharpiegate was constructed with the distinct help of the bottom-up dynamics of the crowd. In August 2020, President Trump declared in his rally in Oshkosh, Wisconsin, that "The only way we're going to lose this election is if the election is rigged".[19] After months of public statements priming his supporters for a rigged election, Trump laid the groundwork upon which the bottom-up nature of the Sharpiegate dynamic could grow. Well-meaning individuals took up this call and began creating casebooks of online 'evidence', detailing false witnesses of "election fraud", which could encompass any instance from an allegedly dead voter casting a ballot, to an incorrect number of mail-in ballots arriving in the mail, or a case of mail-in ballots being dumped in a ditch. Verifying whether any individual vote is an instance of fraud is a relatively personal and straightforward process. However, the political views and prevailing narratives about potential election fraud contributed to these individuals' ongoing misinterpretation of what they were experiencing and motivated them to amplify the content widely.

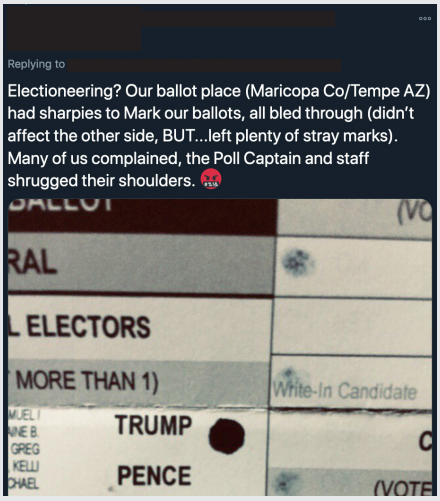
Once introduced onto social media, these cases of false witnesses of "election fraud" were frequently picked up and amplified by influencers and rank-and-file accounts alike. In the case of Sharpiegate, the 'evidence' gathered by the crowds was the cumulative experience of many individuals who had received a Sharpie brand pen to mark their ballot at the polling station.

14 Citron, *Hate Crimes in Cyberspace*.
15 Howard and Hussain, *State Power 2.0*.
16 Romm, Stanley-Becker, and Dwoskin, 'Facebook Bans "STOP THE STEAL" Group Trump Allies Were Using to Organize Protests against Vote Counting'.
17 Sterrett et al., 'Who Shared It?'.
18 Citron, *Hate Crimes in Cyberspace*.
19 Smith, 'Trump Has Longstanding History of Calling Elections "rigged" If He Doesn't like the Results'.

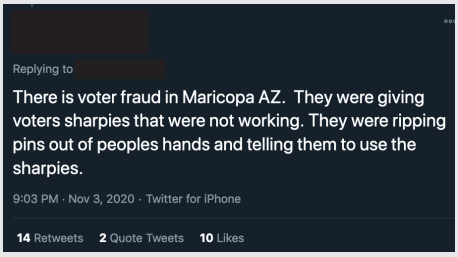
A common misconception that the Sharpie would 'bleed' through the ballot caused many voters to become genuinely concerned about the status of their vote. What began as a common point of concern, however, quickly combined with the context of President Trump's electoral fraud narratives, turning the initially concerned tone into suspicion, then accusations of genuine fraud.



Concern



Suspicion



Accusation

Source: Election Integrity Partnership (2021). 'The Long Fuse: Misinformation and the 2020 Election'.

Once a critical mass of small-scale users began to question the Sharpiegate conspiracy, influential accounts including hyper-partisan media, conservative political figures, and other elite right-wing influencers, assembled this content further to fit the larger 'Big Lie' narratives (e.g., a "rigged election") and spread it to increasingly large audiences. Users understood the power of this echo dynamic and used it to their advantage: if they were able to catch the attention of an influencer through a retweet or a mention, that elite user would promote the content that the original user had gathered and echo it to a new subset of their audience. Subsequently, the relationship between bottom-up and top-down dynamics only tightened, in which both user groups continued to lean into the advantages of the participatory landscape.

Thus, the dynamic of participatory disinformation was completed. This cycle is extremely powerful and resistant to correction, especially in the combination of both top-down and bottom- up sharing patterns. Elites set the agenda and seed the narratives, the audience creates evidence to fit these narratives, refining their narrative to identify what resonates, elites watch this filtering process then pick the most 'valuable' narratives to echo back to their audience, motivating them to find even more evidence to continue this cycle.

Cross-platform coordination

Each platform creates different affordances for information sharing and social interactions. Many of these platforms also incentivize sharing from other social media platforms, as well as the broader digital ecosystem, including personal blogs or news websites. What is more, social media users tend to operate on more than one platform, with their online personas spanning virtual platform boundaries.[20] Thus, information operations will often target users and communities across multiple platforms, adopting a coordinated multi-platform approach that intentionally moves content from one platform to another.

One of the most prominent examples of cross-platform coordination is the *secondary infektion* campaign run by Russian operatives from 2014 to 2020. Researchers found that secondary infektion made more than 2,500 posts across 300 platforms over this six-year time period.[21] While state-run disinformation entities usually hide behind fake profiles or sockpuppet accounts to spread disinformation, influencers take the opposite approach and develop a strong brand across multiple platforms, often linked to a real-world identity.

During the 2020 US presidential election, many conspiratorial narratives were picked up by influencers who used their cross-platform branding to push disinformation across various platforms. Claims about Hammer and Scorecard or the Dominion voting machines started on websites and on Twitter, then spread through YouTube videos, and Parler and Reddit boards. These conversations also drew the attention of hyper-partisan news websites, political influencers, media pundits, and extremist groups, who then re-amplified the narratives back to mainstream and fringe platforms.[22] By moving conversations between platforms, cross-platform information operations limit the effectiveness of any one platform's policy responses.

Algorithmic amplification

Platforms use algorithms – or automated sets of rules or instructions – to transform data into a desired output. Using mathematical formulas, algorithms rate, rank, order and deliver content based on factors such as an individual user's data and personal preferences, or aggregate trends in the interests and behaviour of similar users. The algorithmic curation of content, and the way that news and information are prioritized and curated for users, presents a number of unique challenges for elections and democracy. Do algorithms present diverse viewpoints or reinforce singular ones? Do they nudge users towards extreme or polarizing information? Or do they emphasize sensational, tabloid or junk content over news and other authoritative sources of information?

Although all platforms must grapple with the side effects of algorithmic curation, platforms that are particularly reliant on personalized exploration pages or platform suggestions to drive user action highlight the potential dangers of algorithmic content curation. The TikTok For You page and YouTube's rabbit hole of 'Up Next' recommendations are two salient examples of algorithmic curation in which users interact most heavily with content curated by the platform's algorithm. In contrast to Instagram's newsfeed, which is populated almost exclusively by content from the network a user has *chosen* to follow, the For You page on TikTok is an exploratory feature which surfaces obscure videos that may be completely unrelated to a user's expressed preferences.

Although this algorithmic curation can be largely harmless, this model can quickly adapt to a new user's initial distrust of traditional sources to provide a stream of 'alternative' viewpoints that snowball what may start as a slight deviation from accepted sources into a feed overrun by extreme or conspiratorial viewpoints. This rabbit-hole dynamic has been observed to be particularly strong on TikTok, whose recommendations system is both the key to its success and its biggest vulnerability to new disinformation attacks.[23]

Images and ephemeral content

The ethereal 'story' feature of social media was first pioneered by Snapchat in 2013, and quickly solidified as a core facet of online platform communications. Through this feature, users are able to post content with a longevity between that of

20 Omnicore, '70+ Social Media Statistics You Need to Know in 2021'.
21 Nimmo et al., 'Secondary Infektion'.
22 Election Integrity Partnership, 'The Long Fuse'.
23 Hickey, 'TikTok Played a Key Role in MAGA Radicalization'.

an instant or disappearing message (such as in Snapchat's disappearing image functionality) and a longstanding post. Often, stories are hosted for 24 hours before disappearing, and users are able to host not only their own content but the content of friends on their own stories.

In the summer leading up to the 2020 election, there was an influx of 'political Instagram story activism', especially among younger populations[24]. Suddenly, important content about when, where, and how to vote was casually being shared through Instagram Stories, where the election-related content in question benefited from the trustworthiness of the perceived endorsement of the poster. As Stories are transient features, a user may feel more comfortable posting a half-baked or non-verified claim onto them, which they would generally not post on their timeline on account of its greater longevity. Story content often gives a 'preview' of a longer post. However, as a preview, the longform explanation that accompanies this initial image in the post's caption is lost when the content is pushed to Story functionality, ripping away all possible context. Finally, live 'evidence' is often also featured within the Story functionality, as users document the real-world actions, protests, or hiccups that best align with their framing of the offline situation in short video clips or images. Visual content is a powerful vector to spur offline action, magnifying the importance of the story format for information sharing.

Albeit subtle, the combination of these product features makes Stories a key vulnerability for platforms facing disinformation onslaughts: users use this feature to share content quickly, often without verification, which may represent a more extreme viewpoint due to the transient nature of the content, which is almost always presented without context. Since the act of putting a piece of content into a story is not calculated as a new user engagement with that content (e.g. as a Like or Share), Story views and reshares are almost completely invisible to external researchers attempting to gauge the spread of a particular narrative that has been picked up through this feature.

Live streams

Social media platforms, like YouTube, Facebook and Twitch, have provided users with the ability to stream content – in real time – to large audiences. Livestreamers can have significant reach, with some individuals or channels reaching audiences comparable in size to mainstream news outlets. During the 2020 US presidential election, livestreams became an interesting front for the spread of mis- and disinformation as a variety of conservative influencers, media personalities, and ordinary citizens were able to use livestream features to discuss the election, while sometimes perpetuating claims about social unrest or voter fraud.

Livestreams create complex moderation challenges for platforms wishing to minimize mis- and disinformation, as the streams are often boosted in the moment by platform algorithms, although there is little opportunity to address claims in real time. Similarly, while a specific tweet or Facebook post can easily be identified and removed, livestreams that contain mis- or disinformation are much harder to remove because it is not possible to restrict a small fraction of an extended video. Thus, enforcement action has to be taken against the entire video or channel, or not at all. Finally, livestreams can also be ephemeral, like the other kinds of content described above, and do not always remain online after they have been streamed. This presents challenges for removing accounts that use livestreams as a way to push mis- and disinformation to audiences as there is no public evidence trail.

Extremist platforms

As large, mainstream, and highly viral platforms like Facebook, Twitter and YouTube have drawn criticism for the increasing spread of mis- and disinformation online, social media companies have begun taking action against content and accounts that breach their community guidelines. The removal of accounts and content has provided an opportunity for platforms with a more laissez-faire approach to content moderation to draw audiences from extremist or conspiratorial communities. Indeed, many far-right and conservative audiences have moved to alternative, fringe and extremist communities because of their growing concerns over censorship on mainstream social platforms like Facebook.

Platforms like Gab, Parler, or 4Chan allow extremist views to come together without over-

24 Brennan, 'How Effective is Instagram Story Activism, Really?'.

sight, and there are growing concerns that these spaces contribute to ongoing radicalization. During the 2020 US presidential election, alternative platforms like Parler created an echo chamber for racism and hatred that escalated in the leadup to the vote, ultimately culminating in the insurrection at the Capitol on 6 January 2021. Comments calling for the "rape, torture and the assassination of public named officials and private citizens" were left unmoderated by the platform.[25]

Extremist communities face challenges for content moderation because they do not have the same standards, transparency or accountability for moderation as large, mainstream social media platforms. But this also points to the important role of moderation by other kinds of internet companies, which are less visible than large, public-facing, social media platforms.[26] After the insurrection at the Capitol building, Parler was removed from the Apple App Store and the Google Play Store, which prevented users from downloading the app onto their mobile phones.[27] Amazon Web Services (AWS) also took action against Parler, and discontinued its web hosting services, removing the infrastructure it needed to operate on the internet.[28] Similar actions by internet intermediaries have been taken against 8chan and other extremist communities to limit the spread of harmful information online in the past.[29] These examples serve as a reminder that content moderation happens across all levels of the internet, and that social media platforms are just one actor making decisions about the accessibility of online content.

Encrypted platforms

Direct messages and encrypted chat platforms are the most private channels through which users can share information. Recently, the rise in popularity of platforms such as WhatsApp, Telegram, and Signal has increased the popularity of this means of communication. This is especially true in countries such as Brazil, India, and throughout the Global South, where WhatsApp is one of the primary mediums for communication between people, and a large source of news and information.[30] Chat applications are also incredibly popular among diaspora communities. According to the Pew Research Center, 42% of Hispanics in the United States use WhatsApp, compared to 24% of Black Americans and 13% of white Americans, helping immigrants and diaspora communities to stay in touch with family, friends, and relatives that may live abroad.[31]

Users more often communicate with individuals they know 'offline' in one-to-one contexts, although some platforms like Telegram offer group functions that allow for broadcasting and interaction with a larger network of people. When it comes to one-to-one communications that all encrypted messaging applications afford, they often benefit from a strong feeling of trust between users. Disinformation shared in one-to-one communications can be incredibly powerful because of the trust that users have in, and connection they have with, friends, family and those in their close network.[32] However, information spreads more slowly in one-to-one conversation contexts, as content relies on organic user shares and does not directly benefit from algorithmic amplification or virality dynamics.

Overall, disinformation is difficult to quantify in these one-to-one communication environments, precisely because of the privacy affordances that this dynamic is meant to achieve. While encryption is important for protecting the security and safety of human rights defenders around the world, the purveyors of disinformation have leveraged the security of these platforms to enhance the spread of harmful or misleading information online. Furthermore, even in unencrypted one-to-one platform dynamics, interventions for disinformation such as labelling or takedowns are not feasible as a product feature since users do not expect moderation or labelling in what is perceived to be an 'off-platform' communication channel.

25 Parler LLC v. Amazon Web Services Inc.
26 Bradshaw and DeNardis, 'The Companies More Powerful than Facebook'.
27 Peters, 'Apple Removes Parler from the App Store'.
28 Palmer, 'Amazon Says It's Been Flagging Violent Posts to Parler since November'.
29 Stewart, '8chan, Explained'.
30 Newman et al., 'Reuters Institute Digital News Report 2020'.
31 Perrin and Anderson, 'Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018'.
32 Gursky, Riedl, and Woolley, 'The Disinformation Threat to Diaspora Communities in Encrypted Chat Apps'.

Future technologies: Artificial intelligence, deep fakes and the next generation of disinformation

Technology itself is constantly evolving, and many new technological innovations around artificial intelligence (AI) are creating new affordances for propaganda and persuasion. Automated bot accounts can use machine learning algorithms (like GPT-3) to sound more human; generative adversarial networks (GANs) can be used to create fake profile pictures that look like real people and other forms of synthetic media like "deep fake" videos; and more sensors brought about by the Internet of Things create more data about individuals, their behaviours, and movements, which can be used to better refine targeted messaging and advertisements.

There have been ongoing concerns about the use of AI and other technologies to disrupt elections around the world. In 2019 the Worldwide Threat Assessment suggested that Russian and other American adversaries "probably will attempt to use deepfake or similar machine-learning technologies to create convincing but false image, audio and video files to augment influence campaigns directed against the United States and our allies and partners".[33] Indeed, many of these technologies have already been applied to undermine the credibility of women by depicting them in sexual or pornographic videos[34] and, in some instances, have been used by politicians, as was the case in India when the president of the ruling party (Bharatiya Janata Party [BJP]), Manoj Tiwari, created a deep fake as part of their campaign.[35]

However, during the 2020 US presidential election, many of these technological innovations remained untapped. While there were ongoing concerns that deep fakes and other AI-driven technologies might be used to undermine the election in the leadup to voting day, influence operations made use of many of the existing affordances of platform technologies – such as livestreams, bottom-up group dynamics or amplification via influencers. Thus, policy responses to influence operations and mis- and disinformation on social media must consider contemporary affordances, while tracking innovations in future technologies.

33 Lewis, 'Trust Your Eyes?'.
34 Hao, 'Deepfake Porn Is Ruining Women's Lives. Now the Law May Finally Ban It'.
35 Jee, 'An Indian Politician Is Using Deepfake Technology to Win New Voters'.

Evaluating platform responses

The first line of defence against disinformation is the response of the social media platforms on which these narratives may be hosted. Policymakers preparing to respond to a large-scale disinformation attack should be aware of how platforms are organized against these attacks, in order to both supplement and advocate improvements to this first line of defence.

In 2016, social media platforms were unprepared for the disinformation advent of that year. The story of disinformation in 2016 was predominantly one of foreign interference and inauthentic behaviour. Since then, social media companies have recalibrated their policies and overhauled their community guidelines to more directly address coordinated inauthentic behaviour occurring on their platforms. However, the reactive nature of these policy responses with regard to actors spreading disinformation about politics was largely limited in scope to inauthentic coordination and state-backed operations. In preparation for the US 2020 election, social media companies also introduced specific content-focused 'civic integrity' policies as addenda to their existing guidelines to remove content that sought to delegitimize the election or incite violence at polling stations.

During the US 2020 election, the Election Integrity Partnership (EIP) catalogued different platforms' content-based policies, and gauged their efficacy across key election-related categories of disinformation. Results were varied across platforms, and although the content-based approach did show a marked improvement on the state of policies heading into the 2016 US presidential election, many platforms remained without any election-related policies at all. Moreover, after observing the tactical updates that were debuted during the election period, the EIP team found that platforms' content-focused strategies were insufficient to address the core product affordances that accelerated the spread of the most pervasive disinformation narratives, including #StopTheSteal.

The disinformation landscape has evolved quickly since the days of sockpuppet accounts spreading Twitter memes, and policymakers must demand that platforms act accordingly to develop clear guidelines to confront the dynamics that slipped through their actor- and content-focused frameworks. Existing platform policy challenges can be grouped into five main buckets:

1) The most salient disinformation narratives were **coordinated by real users**, rather than fake, inauthentic accounts.
2) While platforms have policies targeting election disinformation, **not all content was clearly falsifiable** and did not always appear harmful or use the vocabulary of violence.
3) **Disinformation was networked** across many platforms, including unmoderated or encrypted spaces making it hard to track and remove harmful content at scale across the information landscape.
4) Certain narratives about election delegitimization continued to be **algorithmically amplified** by platforms whose business models continue to incentivize the viral spread of disinformation.
5) Future innovations around **AI, deepfake technology and increased data surveillance continue to pose a threat to future disinformation campaigns**, but many of these potentials have remained untapped.

In addition to these gaps, an ongoing struggle for platforms will be the actual enforcement of their policies, as well as the enforcement of any new policies they implement. Platforms' Terms of Service and Community Guidelines create broad rules for the removal, suspension, or demotion of content and accounts, although moderating at scale and in real time presents real challenges, especially when disinformation narratives are not easily falsifiable or not clearly harmful.

Facing coordinated authentic behaviour

The global phenomenon of the top-down and participatory dynamics of disinformation has highlighted the domestic, authentic brand of disinformation as perhaps the most difficult and most pervasive incidence of its spread. Platforms are rigorous in their policies towards coordinated inauthentic behaviour. However, authentic coordination can sometimes look very similar to inauthentic coordination online. Facebook's Inauthentic Behavior Policy specifies that it will only apply in situations where "the use of fake accounts is central to the operation". Twitter similarly specifies its interpretation of coordination as "coordination — creating multiple accounts to post duplicative content or create fake engagement", with additional language highlighting the "inauthentic" requirement needed for this policy to be enacted on a specific account's activities. Almost all major platform policies have veered into highly militarized language that emphasizes the importance of "foreign" speech when discussing "coordinated behaviour", further complicating the extent to which these policies would be applied to authentic and domestic users.

The constraints of this policy language and the precedent set by social media platforms to only act against coordination in the more obvious cases of clear impersonation have created a vulnerability in their willingness to take action against content created by real users, especially those acting domestically, with very similar behavioural patterns to those of coordinated inauthentic networks. These dynamics may arise both in the case of top-down disinformation, as well as within group and community platform spaces, where users can organize themselves towards driving engagement for a hashtag or towards a specific user. Even beyond clear calls to action, verified accounts have been found to engage in 'institutionalized' posting patterns, consistently messaging at the same moments to overwhelm their combined millions of followers and reinforce the desired message in a more nuanced fashion. This dynamic blurs the line between healthy public debate and exploitation of online networks, which are constructed to over-emphasize the role of these influencers and their messaging.

Policy-evading content

Platforms have created the bedrock of their content-specific policies by mirroring the legal interpretations of acceptable speech in the offline world. Hate speech, sexual or exploitative imagery, and even copyright infringement have clear online platform guidelines, sometimes even enforcing user suspensions for posting this content. However, platform policies are less developed with respect to novel vulnerabilities resulting from the distinct features of online conversation. This includes aforementioned dynamics such as the rapid spread of ephemeral image content, platform live streams, and policy-evading content framing strategies.

Stories and live stream content are distinct product features, exemplifying new ways that users can share their ideas and views with a large audience. The real-time nature of these features can make them harder to moderate: both in terms of the processing power needed to moderate video in real time, for example, but also due to how users share and absorb information that is presented in a real-time capacity. Labelling is more difficult for live streams: absent a delay on the video itself, users will have already been exposed to the violative content by the time any sort of real-time acknowledgement is generated and applied to the livestream. In the case of stories, the context accompanying this information-sharing vector is already very limited, which makes labelling difficult, and there is no empirical evidence to suggest that labelling stories will actually work either way. Furthermore, the emergent behavioural dynamics of story material may exacerbate the effects of this information-sharing vector: ephemeral content may be perceived by some users as 'lower stakes', in which they can engage with and broadcast an opinion more extreme than their default stance, since the content will only be associated with their profile for 24hrs rather than forever until deletion. These behaviours and the associated repercussions of these product features require platforms to more strategically moderate these vectors, perhaps enforcing their policies in a stricter way, or by developing novel labelling and information-sharing capacities to correct content shared via these channels.

In the case of policy-evading content framing, platforms must reconsider their relatively strict adherence to offline speech norms, and acknowledge these dynamics as strategic adaptations by disinformation actors to escape moderation. Malign actors are well aware of platform policies and actively manipulate the phasing and delivery of their content to evade moderation: most recently, this was highlighted by a ring of almost 40 QAnon channels that would tactically highlight each other's videos, then promptly delete them within a week or so of publication to avoid YouTube's review turnaround time, and to avoid being taken off the platform. Anti-vaccine activists are also aware of their infringement of policy guidelines and adapt their phrasing (using terms like 'v@xxine' instead of 'vaccine', for example) and content framing accordingly.[36] As previously highlighted, the 'just asking questions' dynamic is increasingly popular with influencers in particular, who understand they must incorporate a simple question mark ('?') after a statement to avoid content takedowns.

Networked disinformation

Platforms are not siloed and disinformation is networked across platform boundaries. Thus, a suitable response will need to be similarly networked across platforms to be successful. Some platforms, such as TikTok, have very specific product features that highlight the networked nature of disinformation. For example, the 'Green Screen' video filter, which allows users to add whatever background picture they like to their TikTok video, let several major narratives or violative posts that were started on 'mainstream' platforms such as Facebook or Twitter continue their lifespan on TikTok as 'background' images to new viral videos, revitalizing an idea that may have already been labelled or taken down in its previous iteration.[37]

Networked disinformation is complicated by extremist and encrypted platforms. Some extremist communities lack any sort of content guidelines: these platforms may serve as testing grounds for identifying the most salient narratives to export to larger platform audiences. Similarly, encrypted platforms can serve as both A/B testing grounds for larger extremist communities to identify narratives worth spreading on large 'mainstream' platforms, or they may serve as a personalization function, in which users copy narratives from larger platforms and personalize them to circulate among close family or friend groups.

Cross-platform coordination by actors must also be taken into account when discussing an improvement to combat networked disinformation. The Election Integrity Partnership found that the most prominent repeat spreaders of disinformation have active accounts across platforms, serving to unify their own messaging across platforms and their audiences. Cross-platform coordination also highlights the importance of other actors in removing harmful content as web hosting companies, app stores and other internet intermediaries play a role in moderating content beyond social media platforms.

Incentivized disinformation

Social media algorithms prioritize personalized content in order to generate engagement and increase the time users spend on their platforms. Subsequently, content that is provocative, shocking, conspiratorial, or highly engaging garners more likes and views, and might spread further and faster than true or factual information. In the leadup to the 2016 US presidential elections, "fake news" and "junk news" stories were shared more than real news stories.[38] In Europe, junk news stories also generated significantly more likes, shares and comments than professionally produced news and information.[39] During the US 2020 presidential election, narratives like #StoptheSteal continued to be picked up and shared with various audiences across social media platforms, despite their best efforts to limit the spread of disinformation online.[40]

One of the problems with analyzing the impact of social media algorithms on the spread of disin-

36 Koltai and Moran, 'Avoiding Content Moderation'.
37 Bradshaw et al., 'Election Delegitimization'.
38 Silverman, 'This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook'.
39 Marchal et al., 'Junk News during the EU Parliamentary Elections'.
40 Silverman, Mac, and Lytvynenko, 'Facebook Knows It Was Used To Help Incite The Capitol Insurrection'.

formation is that academics, journalists, and civil society actors lack access to sufficient data to understand the problem. While algorithms might deliver content to users based on large audience trends, content is also tailored to individuals based on their data and their previous interactions with the platform. Understanding the role of social media algorithms in incentivizing the spread of disinformation, or pushing certain groups of users towards extreme or conspiratorial content is hard to measure without better access to the data held by social media companies. Balancing user privacy with the need for more and better data is an ongoing challenge for policymakers and academics studying the impact of social media on democracy and politics.

AI on the horizon

Innovations in technology will continue to change the nature of disinformation and how it spreads across our information and communication ecosystem. While AI tools like "generative adversarial networks" have been applied to improve digital deception and account sock puppetry, many of these tools have not yet seen widespread deployment in the realm of politics. However, technologists and policymakers should still consider the effect that these technologies can have on disinformation campaigns.

AI-driven technologies, like so-called deep fake videos or the GPT-3 technologies that create real-looking videos or image-based content, might be just as persuasive as real content. This is partially because video and image-based content is more persuasive than text or article-based content. But it is not just about fooling people with fake images. Deep fakes and the increasing use of these deceptive technologies can also lead to an "exhaustion of critical thinking", whereby it takes an increasing amount of energy for users to discern real information from fakery.[41] Some preliminary academic studies have shown that as platforms, fact-checkers, and journalists continue to label mis- and disinformation online, audiences lose trust in the credibility of all information, including news from highly professionalized outlets.[42] This overall loss of trust in the information environment might pose an even greater risk than the deceptive technologies themselves, as technologists and governments continue to invest in technologies that can identify AI-generated or manipulated videos.

41 Engler, 'Fighting Deepfakes When Detection Fails'.
42 Luo, Hancock, and Markowitz, 'Credibility Perceptions and Detection Accuracy of Fake News Headlines on Social Media'.

Conclusions and recommendations

Democratic societies depend on citizens having access to trustworthy information in order to make informed decisions about everyday choices that affect their future wellbeing: from elections, to vaccination decisions, and more. Maintaining a healthy information environment, however, requires collaboration across government entities, social media platforms, civil society groups, and research organizations dedicated to the analysis of online influence operations. The full spectrum of disinformation identification, analysis, and response is a whole-of-society endeavour that cannot be built overnight. Relationships must be fostered early and maintained across a variety of disinformation threats, in order to create a robust response mechanism unique to a specific country's distribution of resources and authority to face these problems.

In order to build the infrastructure to confront this new wave of disinformation threats, policymakers must understand both the state of their current information ecosystems, as well as the potential evolutions in the core dynamics described above. Although the list below is not comprehensive, it covers some core questions regarding the platforms, actors, and communities of interest that must first be assessed:

- What platforms are used by users in your jurisdiction? Are different demographics congregating in different spaces?
- What facets of your core infrastructure require a baseline degree of public trust? Examples may include elections, public health, communications and 5G, climate, and other facets of government infrastructure.
- What are the major ideological communities that drive online conversation? Which are important to your specific disinformation attack? This may include conspiratorial groups such as QAnon, interest groups such as the anti-vaccine community, political groups, specific age demographics, and more.
- For each ideological community, who are the major influencers? What kind of influence do they have? What has their messaging been like?
- What are the major media entities that may affect public opinion? What is their relative size, influence, and ideological bent? Can you identify and begin to track their online assets across different platforms?
- Have any major narratives about your public infrastructure in question already taken hold? What are they? Are they likely to continue? Are there trusted groups who could debunk these narratives?
- What is the state of your own online presence? Do you have a verified account to represent the government-associated entity actually in charge of delivering the public goods in question? Do local officials also have verified accounts?
- Do you have a history of targeting by state actors? If so, are there any residual assets from these foreign influence attacks to be aware of? Do you need to consider foreign state media driving domestic opinion?

Different countries will be susceptible to different tactics, actors, and content dynamics. Therefore, a basic understanding of the actors, communities, platforms, and tactics historically used by disinformation actors in a specific jurisdiction is necessary to understand how subsequent relationships to platform policy teams, civil society organizations, and research institutions are formed.

The US government was caught off guard by the unique challenges of the 2020 election: by not pre-empting how platform policies had evolved to aggressively counter foreign interference run by bot and sockpuppet accounts, government entities found themselves over-prepared for the wrong challenge. The domestic, influencer-driven dynamics of viral misinformation that culminated in the January 6th attack on the US Capitol are far more difficult to counter than a bot-led campaign, as platforms are less likely to take action against

the innocuous presentation of modern disinformation campaigns. An increasingly sceptical and networked public is only becoming more susceptible to the dynamics observed in the US. Hence, democratic institutions must act quickly and collectively to identify the threats to their own populations, and build resilient coalitions to pre-empt and counter such threats.

Authors

Isabella Garcia-Camargo is currently a consultant at the Krebs Stamos Group, where she focuses on disinformation and cybersecurity crisis management for corporate clients. In her most recent role at the Stanford Internet Observatory, Isabella helped to launch the Election Integrity Partnership and later the Virality Project, two coalitions that united government, civil society, social media, and traditional media institutions in the shared goal of real-time disinformation response. Isabella graduated from Stanford's Computer Science department, where she focused on artificial intelligence and ethics in technology.

Samantha Bradshaw is a postdoctoral fellow at the Stanford Internet Observatory and the Digital Civil Society Lab at Stanford University. She is a scholar of technology and politics and writes frequently about the relationship between social media and democracy. Her research has been published in leading academic journals and has been featured by global media outlets, including the New York Times, the Washington Post, CNN, the Globe and Mail, the Financial Times, and Bloomberg Magazine. Samantha completed her DPhil at Oxford University in 2020.

References

Abbas Ahmadi, Ali, and Ester Chan. 'Online Influencers Have Become Powerful Vectors in Promoting False Information and Conspiracy Theories'. First Draft, December 8, 2020. https://firstdraftnews.org:443/latest/influencers-vectors-misinformation/.

Bessi, Alessandro, and Emilio Ferrara. 'Social Bots Distort the 2016 U.S. Presidential Election Online Discussion'. *First Monday*, November 3, 2016. https://doi.org/10.5210/fm.v21i11.7090.

Bradshaw, Samantha, and Laura DeNardis. 'The Companies More Powerful than Facebook'. Centre for International Governance Innovation, Forthcoming.

Bradshaw, Samantha, and Philip N. Howard. 'Global Disinformation Disorder: 2019 Inventory of Social Media Manipulation'. Computational Propaganda Project Working Paper Series. Oxford: Oxford Internet Institute, September 2019. https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/Cyber-Troop-Report19.pdf.

Bradshaw, Samantha, David Thiel, Carly Miller, and Renee DiResta. 'Election Delegitimization: Coming to You Live'. Election Integrity Partnership, November 17, 2021. https://www.eipartnership.net/rapid-response/election-delegitimization-coming-to-you-live.

Brennan, Megan. 'How Effective is Instagram Story Activism, Really?'. Vanderbilt Political Review, 19 October 2020. https://vanderbiltpoliticalreview.com/9659/us/how-effective-is-instagram-story-activism-really/.

Citron, Danielle. *Hate Crimes in Cyberspace*. Cambridge: MIT Press, 2015.

DiResta, Renee, and Matt DeButts. '"Newsworthiness", Trump, and the Facebook Oversight Board'. *Columbia Journalism Review*, April 26, 2021. https://www.cjr.org/the_new_gatekeepers/facebook-oversight-board-2.php.

Election Integrity Partnership. 'Evaluating Platform Election-Related Speech Policies'. Election Integrity Partnership, October 28, 2020. https://www.eipartnership.net/policy-analysis/platform-policies.

Election Integrity Partnership. 'The Long Fuse: Misinformation and the 2020 Election', 2021. https://purl.stanford.edu/tr171zs0069.

Engler, Alex. 'Fighting Deepfakes When Detection Fails'. *Brookings* (blog), November 14, 2019. https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/.

Guadagno, Rosanna E. 'Compliance'. The Oxford Handbook of Social Influence, September 7, 2017. https://doi.org/10.1093/oxfordhb/9780199859870.013.4.

Gursky, Jacob, Martin J. Riedl, and Samuel Woolley. 'The Disinformation Threat to Diaspora Communities in Encrypted Chat Apps'. *Brookings* (blog), March 19, 2021. https://www.brookings.edu/techstream/the-disinformation-threat-to-diaspora-communities-in-encrypted-chat-apps/.

Hao, Karen. 'Deepfake Porn Is Ruining Women's Lives. Now the Law May Finally Ban It'. MIT Technology Review, February 12, 2021. https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/.

Hickey, Cameron. 'TikTok Played a Key Role in MAGA Radicalization'. *Wired*, 2021. https://www.wired.com/story/opinion-tiktok-played-a-key-role-in-maga-radicalization.

Howard, Philip N., and Muzammil M. Hussain, eds. *State Power 2.0: Authoritarian Entrenchment and Political Engagement Worldwide*. Farnham, Surrey; Burlington, Vermont: Ashgate, 2013.

Jee, Charlotte. 'An Indian Politician Is Using Deepfake Technology to Win New Voters'. *MIT Technology Review*, 2019. https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/.

Koltai, Koko, and Rachel Moran. 'Avoiding Content Moderation'. *Virality Project* (blog), Forthcoming.

Koltai, Kolina, Jack Nasseta, and Kate Starbird. '"Friend of a Friend" Stories as a Vehicle for Misinformation'. Election Integrity Partnership, October 26, 2020. https://www.eipartnership.net/rapid-response/friend-of-a-friend-stories-as-a-vehicle-for-misinformation.

Lewis, James. 'Trust Your Eyes? Deepfakes Policy Brief', October 23, 2019. https://www.csis.org/analysis/trust-your-eyes-deepfakes-policy-brief.

Luo, Mufan, Jeffrey T. Hancock, and David M. Markowitz. 'Credibility Perceptions and Detection Accuracy of Fake News Headlines on Social Media: Effects of Truth-Bias and Endorsement Cues'. *Communication Research*, May 23, 2020, 0093650220921321. https://doi.org/10.1177/0093650220921321.

Mahtani, Shibani, and Regine Cabato. 'Why Crafty Internet Trolls in the Philippines May Be Coming to a Website near You'. *Washington Post*, July 26, 2019. https://www.washingtonpost.com/world/asia_pacific/why-crafty-internet-trolls-in-the-philippines-may-be-coming-to-a-website-near-you/2019/07/25/c5d42ee2-5c53-11e9-98d4-844088d135f2_story.html.

Marchal, Nahema, Bence Kollanyi, Lisa-Maria Neudert, and Philip N. Howard. 'Junk News during the EU Parliamentary Elections'. The Computational Propaganda Project, 2019. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/05/EU-Data-Memo.pdf.

Mueller, Robert S. 'Report on the Investigation into Russian Interference in the 2016 Presidential Election'. US Department of Justice, 2019.

Newman, Nic, Richard Fletcher, Anne Schultz, Simge Angi, and Rasmus Kleis-Nielsen. 'Reuters Institute Digital News Report 2020', 2020, 112.

Nimmo, Ben, Camille Francois, C. Shawn Eib, Rodrigo Ferreira, Chris Hernon, and Tim Kostelancik. 'Secondary Infektion'. Graphika, 2020. https://secondaryinfektion.org/report/secondary-infektion-at-a-glance/.

Omnicore. '70+ Social Media Statistics You Need to Know in 2021', 2021. http://www.omnicoreagency.com/social-media-statistics/.

Palmer, Annie. 'Amazon Says It's Been Flagging Violent Posts to Parler since November'. CNBC, January 13, 2021. https://www.cnbc.com/2021/01/13/amazon-says-violent-posts-forced-it-to-drop-parler-from-its-web-hosting-service.html.

Parler LLC v. Amazon Web Services Inc. (United States District Court for the Western District of Washington at Seattle January 21, 2021).

Paul, Kari. 'TikTok: False Posts about US Election Reach Hundreds of Thousands'. *The Guardian*, November 6, 2020, sec. Technology. http://www.theguardian.com/technology/2020/nov/05/tiktok-us-election-misinformation.

Perrin, Andrew, and Monica Anderson. 'Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018'. *Pew Research Center* (blog), April 10, 2019. https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/.

Peters, Jay. 'Apple Removes Parler from the App Store'. *The Verge*, January 9, 2021. https://www.theverge.com/2021/1/9/22221730/apple-removes-suspends-bans-parler-app-store.

Robinson, Olga, and Marianna Spring. 'Coronavirus: How Bad Information Goes Viral'. *BBC News*, March 19, 2020. https://www.bbc.com/news/blogs-trending-51931394.

Romm, Tony, Isaac Stanley-Becker, and Elizabeth Dwoskin. 'Facebook Bans "STOP THE STEAL" Group Trump Allies Were Using to Organize Protests against Vote Counting'. *Washington Post*, November 5, 2020. https://www.washingtonpost.com/technology/2020/11/05/facebook-trump-protests/.

Sabbagh, Dan. 'Trump 2016 Campaign "Targeted 3.5m Black Americans to Deter Them from Voting"'. *The Guardian*, September 28, 2020, sec. US news. http://www.theguardian.com/us-news/2020/sep/28/trump-2016-campaign-targeted-35m-black-americans-to-deter-them-from-voting.

Silverman, Craig. 'This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook'. BuzzFeed News, November 16, 2016. https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook.

Silverman, Craig, Ryan Mac, and Jane Lytvynenko. 'Facebook Knows It Was Used To Help Incite The Capitol Insurrection'. *BuzzFeed News*, January 22, 2021. https://www.buzzfeednews.com/article/craigsilverman/facebook-failed-stop-the-steal-insurrection.

Smith, Terrance. 'Trump Has Longstanding History of Calling Elections "rigged" If He Doesn't like the Results'. *ABC News*, November 11, 2021. https://abcnews.go.com/Politics/trump-longstanding-history-calling-elections-rigged-doesnt-results/story?id=74126926.

Starbird, Kate. 'What Is Participatory Disinformation?' Center for an Informed Public, May 26, 2021. https://www.cip.uw.edu/2021/05/26/participatory-disinformation-kate-starbird/.

Sterrett, David, Dan Malato, Jennifer Benz, Liz Kantor, Trevor Tompson, Tom Rosenstiel, Jeff Sonderman, and Kevin Loker. 'Who Shared It?: Deciding What News to Trust on Social Media'. *Digital Journalism* 7, no. 6 (July 3, 2019): 783–801. https://doi.org/10.1080/21670811.2019.1623702.

Stewart, Emily. '8chan, Explained'. Vox, May 3, 2019. https://www.vox.com/recode/2019/5/3/18527214/8chan-walmart-el-paso-shooting-cloudflare-white-nationalism.

Taylor, Emily, Stacie Walsh, and Samantha Bradshaw. 'Industry Responses to the Malicious Use of Social Media'. NATO STRATCOM COE, 2018.

Villasenor, John. 'How to Deal with AI-Enabled Disinformation'. *Brookings* (blog), November 23, 2020. https://www.brookings.edu/research/how-to-deal-with-ai-enabled-disinformation/.


Hybrid CoE