Tactics, Threats & Targets: Modeling Disinformation and its Mitigation

Shujaat Mirza
New York University
shujaat.mirza@nyu.edu

Labeeba Begum
New York University Abu Dhabi
labeeba@nyu.edu

Liang Niu
New York University
liang.niu@nyu.edu

Sarah Pardo
New York University Abu Dhabi
sarah.pardo@nyu.edu

Azza Abouzied
New York University Abu Dhabi
azza@nyu.edu

Paolo Papotti
EURECOM
papotti@eurecom.fr

Christina Pöpper
New York University Abu Dhabi
christina.poepper@nyu.edu

***Abstract*—Disinformation can be used to sway public opinion toward a certain political or economic direction, adversely impact public health, and mobilize groups to engage in violent disobedience. A major challenge in mitigation is scarcity: disinformation is widespread but its mitigators are few. In this work, we interview fact-checkers, journalists, trust and safety specialists, researchers, and analysts who work in different organizations tackling problematic information across the world. From this interview study, we develop an understanding of the reality of combating disinformation across domains, and we use our findings to derive a cybersecurity-inspired framework to characterize the threat of disinformation. While related work has developed similar frameworks for conducting analyses and assessment, our work is distinct in providing the means to thoroughly consider the attacker side, their tactics and approaches. We demonstrate the applicability of our framework on several examples of recent disinformation campaigns.**

I. Introduction

> *"We are now increasingly seeing that [disinformation] is seen as a cyber-threat, and certain approaches that we've been taking to tackle cybersecurity issues might be used for disinformation as well. We're seeing quite a lot of overlap starting to emerge between these two areas" —Kai (P11)*

Billions of people use online media to consume news and communicate. The current digital landscape facilitates high volumes of information, but promotes low levels of scrutiny by those who consume it, degrading the quality of information in circulation and opening the potential for abuse by targeted attacks. For example, disinformation campaigns were used to sway the British public to vote for Brexit [86]; disinformation on the integrity of the US 2020 elections incited an armed mob, leading to loss of life [85]; and anti-vaccination campaigns have led to Measles outbreaks [11] and are potentially prolonging the Covid-19 crisis [29].

Misinformation and its motivated counterpart, disinformation, pose an increasing threat to society: democratic processes, public safety, and commercial systems are at risk. Advances in technology, combined with the sheer pervasiveness of digital media outlets, have spread the ability to manipulate beyond few highly skilled actors. State and non-state actors alike use online platforms to manufacture consensus, program public opinion in a chosen direction, automate ideological suppression, and undermine civil rights.

Researchers and practitioners have called for the designation of coordinated disinformation campaigns as a cybersecurity concern, given the significant overlap between the two in terms of tools and methods of attack [12], [23]. Disinformation campaigns share a common structure with classic cybersecurity threats: in an adversarial situation, a motivated agent threatens their victim through digital means, often across a network and in a distributed fashion. However, the application of cybersecurity *frameworks* to understand the disinformation landscape and mitigation is still largely unexplored. We propose to bridge this gap by applying security threat modeling to the threat of disinformation. Systematically characterizing an attacker's profile, their likely attack patterns, their most-desired targets, and their commonly-deployed techniques can empower disinformation mitigators to effectively tackle dynamic threats under limited resources.

In this work, we develop a cybersecurity-inspired framework for analyzing disinformation threats. To ground our model in an understanding of the day-to-day reality of the fight against disinformation, we conducted a series of expert interviews (n = 22) with mis-/disinformation mitigators whose experience and training ranged from fact checking and journalism, to platform trust and safety, to conducting research in academia, industry, or NGOs. These inside accounts provide a diverse, practical coverage of the current state of disinformation, and also reveal the priorities and mitigation strategies deployed in the field. Through qualitative data analysis, we identify patterns in the workflows of these experts, uncovering criteria and approaches for the detection, assessment, and mitigation of disinformation operations. We then perform a detailed characterization of threats situated in this landscape, by systematically defining threat actors, their likely targets, their attack patterns, and their attack channels.

We find from the interviews that in practice, mitigators are often unable to operate by a structured method of evaluating the severity of disinformation threats, and they lack formal models or measures to guide their decisions. Our interviews

Network and Distributed System Security (NDSS) Symposium 2023
27 February - 3 March 2023, San Diego, CA, USA
ISBN 1-891562-83-5
https://dx.doi.org/10.14722/ndss.2023.23657
www.ndss-symposium.org

revealed a consistent desire among experts for more-structured approaches to the problem they face, and their accounts of their workflows suggested that they can benefit from a systematic framework. Their first-person accounts support the idea that a security-inspired framework of threat actors, attack patterns, channels, and target audiences can strengthen their fight against disinformation.

The key contributions of our work are as follows:

1) We provide in-depth insight into the work and practices of a diverse group of mis-/disinformation mitigators, extracting their functions and workflow patterns (Sec. IV-A-IV-B) and identifying challenges to their ability to effectively mitigate threats.
2) We apply security threat modeling practices to the disinformation landscape (Sec. IV-C), with insights directly informed by the experience of mitigation experts. We connect our empirical findings to threat characterization practices in security literature. To the best of our knowledge, our study is the first to take this approach.
3) We demonstrate the usefulness of our disinformation threat framework by applying it to recent disinformation campaigns (Sec. V-A). We find that the framework may be a foundation for developing a disinformation threat scoring system, which could eventually support practitioners in their mitigation efforts (Sec. V-B).

II. Preliminaries

A. Terminology

Many works have developed taxonomies and definitions for the misinformation and disinformation space [26], [47], [52], [114], [115]. In this work, we use the term *misinformation* to describe false or incomplete information which is generated or spread by a person who believes it to be true [129]. *Misinformation* implies the absence of intention to mislead. We use *disinformation* to refer to the deliberate dissemination of false information, with the intent to mislead [53], [56]. A *misinformation incident* is a single occurrence of a piece of misinformation. A *disinformation campaign* or *operation* is a coordinated effort by individuals or groups to manipulate public opinion and change how people perceive events in the world, by intentionally producing or amplifying misinformation [80], [96]. A *disinformation campaign* may be comprised of multiple *misinformation incidents* over time.

We use *mitigators* to refer collectively to our participants, professionals such as fact checkers, researchers, trust and safety specialists, whose work focuses on the mitigation of misinformation incidents and disinformation campaigns. Although it sometimes differs from our definitions, throughout the paper we preserve the exact quotations from interview transcripts to retain participants' individual usage of terms.

B. Threat Modeling

A central aspect of cybersecurity is the development and use of threat modeling methods [91]. Threat models abstract a critical system to identify its vulnerabilities, develop profiles of possible attackers, and build a catalog of potential attacks. Security professionals use such models to build defense mechanisms and response protocols. Many industry standards have been specified to assist security professionals and researchers with enumerating attack patterns [101], decomposing attack patterns into tactics and techniques [103], describing the stages of an attack [59], developing robust security programs for organizations [70], and aggregating known weaknesses [102]. Other frameworks provide a serialization format for threat-related objects [71], and a vocabulary for incident characterization and information sharing [108].

C. Related Work

Prior work studying modern disinformation campaigns on online platforms can be organized broadly as focusing on detection [41], [44], [45], [62], [110], [128], assessment [5], [6], [10], [27], [35], [89], [97], [122], [124], and mitigation [3], [31], [51], [77], [81], [90], [99], [126]. In our work, based on expert interviews, we observe a unified pattern in the way that mitigators put these functions into practice.

Wardle et al. [114] suggest that disinformation is defined by its intent to harm. This has inspired treatments of the problem as a type of information warfare [16], [20], [123], [124]; Scheuerman et al. [87] have proposed a framework for characterizing the severity of harm. Our framework complements this perspective by contributing a system for describing the information attacks which lead to harm. Major online platforms recognize the influence of disinformation campaigns on their networks, and approach mitigation by removing content and accounts, then publishing reports on the operations.[1] Modern disinformation campaigns are often conducted across multiple platforms at once [118], and prior work has investigated the ways in which cross-platform attacks can be particularly effective at misleading [57]. Our framework uses a platform-agnostic approach to allow for unified characterization of cross-platform activity within a single campaign.

Disinformation is a global phenomenon, taking on different forms and patterns in different parts of the world. Prior work has studied comparative cases of misinformation in places such as Brazil and India, for instance highlighting actors' choice of different platforms according to regional popularity [50], [95], [29], [30], and distinct regional patterns of biased or toxic speech behavior [37], [38], [84]. Some works develop and demonstrate cross-cultural datasets [79] and tools [66]. Campaigns in different cultures share abstract properties; for instance, every disinformation campaign must have an actor behind it. Our framework offers a standardized taxonomy which can help to highlight shared high-level properties, as well as distinctions in the mechanics of how campaigns are realized, allowing for systematic comparison.

Disinformation is increasingly viewed as a type of cybersecurity threat [12], and prior work has drawn methods from information security intervention to test the use of warning labels for online disinformation [51], [121]. Researchers have also used a security point of view to study risks associated with the use of neural content generation models to produce misinformation [126]. However, ours is the first to take inspiration from security threat modeling for developing a rigorous threat framework to describe and understand disinformation threats.

[1] Facebook calls coordinated campaigns that seek to manipulate public debate *coordinated inauthentic behavior* [68]; Twitter refers to potential foreign campaigns as *information operations* [105].

While there have been some initial proposals for information security tools, such as MISP[2], most open source intelligence (OSINT) systems lack a formal modeling of the disinformation operations. Existing frameworks have focused on points of view of particular stakeholders (e.g., the European Union [74]), studied content beyond disinformation (e.g., harmful content [87]), applied a sociotechnical analysis drawn from computer-supported collaborative work (CSCW) theories (e.g., [96], [127]) or a joint social science and data science lens on vulnerabilities of sociotechnical systems (e.g., Media Manipulation [40]). None of these has used the cybersecurity perspective for characterizing the threats, targets, tactics and channels of disinformation campaigns. Additionally, our work includes insights gathered from a diverse set of experts, and is validated through application on a set of various case studies. Our framework takes steps toward standardizing the practice of modeling disinformation campaigns, so that mitigators can better capture current and future threats.

III. Research Methods

Our research goal is to elucidate the characteristics of disinformation campaigns with a comprehensive view from both the defensive and offensive perspectives. We formulate two research questions, on each side of the problem: *disinformation attack* and *mitigation response*.

1) **Attack:** What characterizes the threat, actors, and severity of disinformation campaigns?
2) **Mitigation:** What characterizes the work, approaches, and operations of disinformation mitigators?

To answer these questions, we conducted open-ended conversational interviews[3] with mitigation and mis-/disinformation research experts. We chose an interview study because it allowed us to access direct insight from a diverse set of experts working in a variety of organizations (industry, academia, NGOs, non-profits), on different areas (e.g., national security, public health), and with different regional focuses. Interviewing experts with a broad set of experiences also ensures that our findings can be generalized across the disinformation landscape. We describe our methods for conducting and analyzing the interviews.

A. Recruiting Participants

We used connections and snowball sampling to recruit mis-/disinformation experts [64]. We initiated our sampling process from participants in a wide range of domains and roles to ensure sufficient coverage. We invited these contacts to voluntarily participate in an unpaid 30-40 minute interview on the topic of disinformation threats. After each interview, we asked the participants to suggest other practitioners with possibly different types of role or organization. To further ensure diversity in our participant pool, we used findings from the interviews to pursue areas which required further exploration by recruiting experts from those areas. For example, based on findings from an initial round of interviews with fact-checkers and journalists, we focused the next round of recruitment on platform trust and safety experts. After conducting 18 interviews, we observed repetition in themes found in subsequent interviews, which we take as an indication of theme saturation given that our recruitment procedure selected for diverse coverage [36]. All interviews took place between July and November 2021.

Table I contains demographic information on our participants. We interviewed 22 experts from 19 different organizations headquartered in different global locations (at most two participants from the same organization). Our participants represent a diverse range of roles (trust and safety specialists (n = 6), fact-checkers (n = 3), academic researchers (n = 3), ...), domains (national security (n = 19), democracy (n = 19), economy (n = 8), public safety (n = 11), and public health (n = 13)), and organization types (media and journalism, academia, NGOs, AI technology companies, and large social media platforms). Disinformation is a multidisciplinary and multifaceted problem, and we selected this variety of roles to understand the different approaches, capabilities, and limitations of practitioners who engage with disinformation in different contexts. Most participants have at least five years of experience in mis-/disinformation mitigation and research (n = 17).

B. Interview Process

For our semi-structured interviews, we developed a slide deck of questions organized around the following main themes: (1) participant background (e.g., role, team, organization, projects); (2) criteria used in surfacing, prioritizing and assessing disinformation projects (e.g., workflows involved, factors observed); (3) characterization of threat actors involved in disinformation campaigns (e.g., attribution, capabilities); and (4) challenges experienced in the process as well as a wish list of tools that could assist them in their jobs (e.g., completeness, usefulness, practicality of different sub-metrics).

All interviews were conducted on Zoom with the slide deck visible to the participants to help direct the conversation. An abridged version of the slide deck content used in the interviews can be found in Appendix A. Multiple authors were present at each interview, but only one of them acted as the main lead for each interview. The others observed silently with the opportunity to propose follow-up questions to the interview lead via direct message.

Before starting an interview, the participants were informed of the goal of the study and their rights as participants. We also obtained their verbal consent to audio-record the interview. We de-identified the participants to protect their anonymity and confidentiality. The audio recordings were transcribed automatically by a transcription software, and these transcriptions were manually corrected by the authors who had attended the live interview before undergoing further analysis. The interviews lasted from 30 minutes up to 1 hour. As we interviewed participants, we refined questions, introduced new questions into the deck which were frequently asked as follow-ups, and modified topics or themes to help better direct the conversation.

C. Qualitative Coding Process

The findings we discuss are the result of systematically organizing our participants' perspectives into an interpretive,

[2]https://www.misp-project.org/

[3]The study received exempt-approval by the Institutional Review Board (IRB) office of the authors' university.

Participant	Role	National Security	Democracy	Economy	Public Safety	Public Health	Years of Experience	Team/Organization Role	Org. Size	Org. Type	Regional Focus
		Domains of Interest									
Hea (P8)	Professor	●	●				6 - 10	Research	51 - 100	Academia	Global
Sam (P19)	Professor	●	●	●	●	●	10+	Research	6 - 10	Academia	Canada, UK, USA
Tay (P20)	Researcher	●	●	●	●	●	6 - 10	Research	6 - 10	Academia	Global
Alex (P1)	Fact-checker	●	●				3 - 5	Fact Checking	6 - 10	Industry	Italy
Babu (P2)	Researcher	●	●	●	●	●	10+	Social Network Analysis	51 - 100	Industry	Global
Dany (P4)	AI-Tech Founder	●	●		●	●	3 - 5	AI Technology Development	100+	Industry	India, UK, USA
Ehan (P5)	Intelligence Analyst	●	●	●	●	●	3 - 5	Social Network Analysis	51 - 100	Industry	Global
Ines (P9)	Fact-checker	●	●	●	●	●	6 - 10	Journalism	11 - 20	Industry	France
Jamie (P10)	Editor	●	●	●	●	●	6 - 10	Journalism	21 - 50	Industry	Global
Lak (P12)	Consultant	●	●				6 - 10	Platform Trust & Safety	6 - 10	Industry	Global
Noel (P14)	AI-Tech Founder					●	10+	AI Technology Development	6 - 10	Industry	Global
Vera (P22)	Data Analyst				●		3 - 5	Outsourced Trust & Safety	11 - 20	Industry	Global
Omar (P15)	Intelligence Analyst	●	●		●		10+	Outsourced Trust & Safety	100+	Industry	Global
Rosa (P18)	Data Scientist	●	●			●	3 - 5	Platform Trust & Safety	21 - 50	Industry	Global
Udo (P21)	Product Manager				●		3 - 5	Outsourced Trust & Safety	11 - 20	Industry	Global
Chan (P3)	Researcher	●	●			●	1 - 2	Research; Advocacy	11 - 20	NGO	Europe
Finn (P6)	Researcher	●	●			●	3 - 5	Research; Advocacy	1 - 5	NGO	Europe
Kai (P11)	Consultant	●	●				6 - 10	Advocacy	11 - 20	NGO	Global
Marge (P13)	Researcher	●	●	●	●	●	6 - 10	Platform Trust & Safety	11 - 20	NGO	Global
Pan (P16)	Researcher	●	●				10+	Think Tank	6 - 10	NGO	Italy
Gada (P7)	Fact-checker	●	●			●	6 - 10	Fact Checking	11 - 20	Non-Profit	Global
Quin (P17)	Researcher	●	●	●			3 - 5	Advocacy; Research	21 - 50	Non-Profit	Global

TABLE I: Participants in our study. We use pseudonyms to protect the participants' anonymity. '●' indicates that a participant mentioned their or their team's expertise in mitigating or researching disinformation within the corresponding domain. Outsourced Trust & Safety are companies that provide trust & safety as a service to other platforms.

analytical framework. We followed an *iterative qualitative coding process* with phases of familiarization (listening to the interviews, reading the transcripts and recording initial impressions or thoughts), open-coding (labeling transcript segments with codes), analytical memo-writing, framework-development (building themes and higher-level categories from the codes), and finally indexing (applying existing categories and codes to the transcripts).

To extract patterns from the interviews in order to develop our threat model, four co-authors reviewed the interviews independently and open-coded a selection of the interviews and created memos. They then compared their codes to find common themes and derive a set of anchoring concepts (actors, tactics, domains, etc.). This was followed by another round of independent coding before a consolidation meeting with all authors. The process resulted in a refined code and category structure that was used to index all the interview transcripts. Our paper reflects the final analytical framework and the findings of this qualitative analysis.

D. Limitations

While we carefully recruited participants with a diverse set of experiences and roles, and from a broad range of organizations in different regions, certain segments are missing, such as experts in cyber-policing agencies. While many of our participants have experience with campaigns conducted across the world, most of them work for US-based or European organizations, and all are based in Global North countries. Not all the experts we attempted to recruit agreed to participate in our study. The study captures disinformation solely from the perspective of mitigators and not the actors. It reflects the views of the experts we interviewed as interpreted by our qualitative analysis. Future work may pursue ethnographic and other observational approaches, or quantitative surveys to corroborate our findings.

IV. Findings

The content of our interviews revealed structures in the work of disinformation mitigation, which we use to develop our disinformation threat framework. We first present findings which provide context for the framework and orientation in the current disinformation landscape. Based on the interviews, we identify *domains* of disinformation work (Section IV-A) and a common pattern in the specific *functions* performed by mitigators (Section IV-B). Building upon these insights, we propose a cybersecurity-inspired threat model to characterize disinformation attacks (Section IV-C). Table III provides a summary of key attack patterns in the model.

A. Domains of Interest

Based on the content of our interviews, we find that there are distinct disinformation *domains*, topics or disciplines where

mitigators focus their work. Given that these reports come from diverse participants, we take direction from their areas of focus to identify five primary domains where the contest between mitigation teams and disinformation actors takes place.

1) National Security: National security includes international relations and conflicts between states. Disinformation attacks on national security have great potential for harm, often supplementing traditional warfare [Omar-P15]. Most participants (n = 19) engage in work related to this domain. Omar (P15)'s investigation into the recent conflict between Armenia and Azerbaijan found that *"domestic Armenian elements, and some backed by Russia, employed significant, heavy disinformation influence campaigns to try to force out the incumbent government."* Kai (P11) explains that in their experience, disinformation can become *"a hindrance to figuring out peace processes or international solutions to a conflict."* Disinformation can also impact conflict situations by altering opinions of other countries or regimes: according to Omar (P15), *"the Iranians will take outspoken, far left academics and they will co-opt them, ... to promote misinformation that has nothing to do with liberals, [such as] the Assad regime in Syria."*

2) Democracy: Many participants (n = 19) focus on disinformation targeting democratic processes such as elections, censuses, referenda, and ballot initiatives. Elections are the most prominent example of a targeted process: actors may seek to directly alter the outcome of the election, or undermine public belief in the fairness of the election. For instance, Hea (P8) describes a project on US election integrity, where they studied mis-/disinformation which questioned the validity of the voting process or caused confusion about when or where to vote. Babu (P2) explains that protecting *"the integrity of the online discourse around the elections"* is of great importance: violating this integrity has potential for *"real harm, impact, or influence"* [Ehan-P5]. Some participants proactively monitor major elections in large, globally powerful states (n = 6) as they are likely targets for disinformation campaigns. Participants (n = 2) also monitor both domestic political groups and foreign states to detect interference in elections.

3) Economy: Disinformation can target financial interests to disrupt market activity, or abuse the financial incentives of platforms to make a profit, and participants (n = 8) work on projects which focus on this domain. For example, in fall 2021, a fake press release stated that Walmart would accept Litecoin for payments, and according to Noel (P14), *"it impacted the stock market because the Litecoin stock went up 32% in 30 minutes ... It looked like it was a real announcement from Walmart and, obviously, that had a big impact on the Litecoin cryptocurrency price."* Disinformation campaigns also take advantage of the monetization schemes of platforms. According to Tay (P20), *"some partisan and false information that we see coming from non-state actors overseas is primarily capitalizing on advertising revenue, particularly thinking about how US advertising revenue is the most profitable."* Rosa (P18) says of their platform, *"the vast majority of violating content is crypto spam or people trying to sell a product or make money."*

4) Public Safety: Some of our participants (n = 11) investigate disinformation campaigns that aim to cause civil unrest or violence. Disinformation narratives often use hate speech to target vulnerable groups and potentially incite hate crimes, so participants (n = 6) monitor hate speech to prioritize their work. Ehan (P5) explains that they investigate suspicious outlets generating content with *"homophobic, Islamophobic, anti-Semitic slurs,"* and Omar (P15) reports that they focus on campaigns in India to address issues of *"communal violence and racism."* These campaigns may cause offline harm to the people they target: Quin (P17) says of their investigation on a campaign which incited violence against a pride march in Georgia, *"it was the day that we saw how online disinformation and calls for violence went offline."* Other threats to public safety occur around crisis events such as climate change (n = 4), natural disasters (n = 2), and man-made disasters (n = 2).

5) Public Health: Participants (n = 13) focus on public health as another high-stakes domain increasingly threatened by disinformation. Health has not always been recognized as a critical domain: Gada (P7) says that in 2019, they experienced frustration with funding priorities in which *"everybody [focused] on political disinformation"* at the expense of investigating *"the biggest problem, of health and science misinformation."* However, the Covid-19 pandemic has reinforced awareness of public health as a critical domain. Chan (P3) explains, *"pretty much everything right now that calls for attention revolves around Covid-19."* Covid-19 misinformation was discussed by most participants (n = 18), and many (n = 12) named Covid anti-vaccination content in particular as a serious concern: Udo (P21) has encountered projects which focus on *"how conversation online would impact or cause harm on the successful rollout of vaccines."*

B. Functions

> *"An analyst turns up to work, the first two hours of the day they spend figuring out ... what am I looking at, what's the fire of the day, the next few hours they try and find more context around it, the next few hours they figure out what should we do about it, and then they report it to a platform, the platform [will] re-verify that independently, and that in turn ... ends up taking 12 hours at best or 24 hours or more, and in the digital world the content is already gone viral, the harm is done and all anyone's doing at that stage is clean up." —Dany (P4)*

While our participants have different roles, areas of focus, and goals, we can largely classify the functions they engage in on a daily basis when working with disinformation into (i) *detection*, searching for potential incidents of interest; (ii) *analysis* of incidents, actors, or networks, often with the goal of contextualizing or evaluating the threat; and (iii) *mitigation*, taking corrective actions to reduce its threat. These functions reflect different stages of a misinformation incident life-cycle and form an integral part of neutralizing disinformation threats. Participants engage in different components of these functions, often serially in a workflow. Various commonly-used tools are summarized in Table II.

1) Detection: In our interviews, we observe two approaches to the detection of disinformation events. The first is a directed approach, where our participants monitor different information feeds, such as tweets, Facebook posts, TV, and news websites, for *known* indicators of disinformation. Some participants maintain lists of known disinformation actors, and monitor feeds for their activity; directed detection can be *"as simple as following as many known malicious actors, or ... known disinformers, across as many networks as possible"* [Tay-P20]. Some participants monitor feeds from

Function	Tool	Count	Use Case
Detection	Botometer	3	Detect bot accounts
	Community leads	2	Flag content (crowd-sourcing)
	Unnamed paid tools	2	Detect violating content
	Twitter trending topics	1	Surface trending content
Analysis	Internal tools/dashboards	6	In-house methods for analysis
	InVid-WeVerify verification plugin	2	Verify content veracity
	Fact-checks (by International Fact Checking Network)	2	Identify narrative trends and actors
	Meltwater	2	Obtain content statistics
	BuzzSumo	1	Obtain content statistics
	ClaimReview	1	Tag fact-checks
	Disinfodex	1	Historical research
	Info. Operations Archive	1	Historical research
	Trendalyzer	1	Visualize information
Detection & Analysis	CrowdTangle (Facebook)	10	Social monitoring platform
	TweetDeck (Twitter)	2	Social media dashboard
	TweetBeaver (Twitter)	1	Data extraction from Twitter
	Birdwatch (Twitter)	1	Community-driven flagging
	tgstat (Telegram)	1	Telegram analytics
	4plebs (4chan)	1	4chan archives

TABLE II: Tools used by the study participants. Count is the frequency of mention by individual participants.

individuals whose activity reaches and influences large audiences, tracking politicians (n = 5), celebrities (n = 2), or political parties and governments (n = 3). Participants also use content-specific identification triggers. For instance, Rosa (P18) searches for particular hashtags and emojis in users' bios, because these signals can indicate QAnon affiliation, and specific categories of content such as Covid misinformation, spam, hate speech, electoral misinformation. A variety of tools are in use to pull feeds from different platforms: Facebook's CrowdTangle (n = 10) has page and account monitoring and tracking features, while TweetDeck (n = 2) and TweetBeaver (n = 1) are used to extract Twitter feeds.

The second is an undirected approach, in which participants monitor information feeds to identify new or emerging incidents for which there may be *no known* indicators. This approach is characterized by dynamic methods which monitor fluctuating activity for anomalies. For example, they may monitor trending topics (n = 4), or content related to breaking news and crisis events (n = 4). Participants also use tools like Botometer (n = 3) to detect anomalous behavior which is likely conducted by automated procedures ("bots"). Some participants use computational methods such as similarity scores to identify the spread of suspicious content (n = 3). In other cases, participants may simply put out a call for tips via Twitter [Tay-P20] or a designated hotline on WhatsApp where people can report misinformation [Alex-P1]. One important reason for undirected monitoring is to cover as many potential blind spots as possible, especially on platforms which are less-studied, or when the resources or expertise of the mitigator is limited. Tay (P20) explains,

> *"if somebody doesn't know how to search through 4chan, they're not going to know that the coordinated campaign started on 4chan or if somebody doesn't have the time or capability to look through hour-long YouTube videos, they're not going to know that a key YouTube influencer amplified that campaign to an audience of millions." —Tay (P20)*

4plebs can be used to monitor activity on 4chan; monitoring YouTube, however, is primarily left to manual review and as yet has limited tools available.

2) Analysis: Analysis can include *contextualization*, where participants connect a specific misinformation incident to its surrounding context. This may include background information on the associated actors, the historical, regional, political, social or cultural backdrop, the overarching narrative or underlying motives, and the historical evolution of the campaign. Ehan (P5) emphasizes the need to acquire *"some basic understanding and knowledge of the region, like the sociopolitical context, the ethnic context."* To this end, one of their *"first steps when ... doing a project is basically to do as much reading as I can on the country or on the region, so that I know I'm not going to be either biased or say something wrong."* Participants also use methods for retrieving the context of the content, to better understand its provenance; for this task, Finn (P6) uses InVid-WeVerify's verification plugin for fact-check lookups and reverse image search. Another cross-platform tool for contextualization is Meltwater (n = 2), which can *"analyze the spread of words to determine who was the first publisher, who was the first one to use a hashtag, what is the coverage around the world, what is the interaction"* [Quin-P17].

Another form of analysis is *activity tracing*, where participants augment their knowledge of an incident with metrics such as shares, to indicate the rate and extent of spread (n = 5), and like or view counts, which can indicate levels of engagement or interaction (n = 5). For example, subsequent to detection, Rosa (P18) conducts a social graph analysis to determine which accounts interact with detected content, *"looking at these profiles and then taking a step up or out [to see] who are all the accounts that they interact with on the platform, is this also an account affiliated with this group?"* Participants often perform tracing with platform-specific tools for Facebook (n = 12) and Twitter (n = 4). Some participants (n = 6) develop their own tools, such as Python scripts which retrieve and visualize these metrics to assess *"the size of this event, how far is it spreading, is it taking off or is it slowing down, what are the main websites, the main platforms, the main influencers [and] which domains are involved"* [Hea-P8].

Analysis can also include *knowledge discovery*, where participants, often researchers, examine a campaign to uncover patterns and behaviors that further our understanding of disinformation and its actors. Hea (P8) explains that while they use specific triggers to identify a lead, once it is identified, their focus shifts to the bigger picture:

> *"We're no longer interested necessarily in what are the precise claims, but how are these claims taking shape, how are they spreading, how are they being countered, is that working, how could that work better ... we look at it on a case by case basis and each case has its own context and its own content, different narratives... [but] what we're really looking at is to try to find some of the commonalities across these cases, so we can start thinking more systematically about solutions." —Hea (P8)*

Knowledge discovery can be assisted by historical repositories such as Disinfodex (n = 1), Information Operations Archive (n = 1), and fact checks published by the International Fact Checking Network (IFCN) (n = 2). For example, Chan (P3) describes their use of the IFCN dataset to retrieve a set of claims about hydroxychloroquine: tracing their origin to Facebook pages with thousands or millions of followers, uncovering a larger pattern, and revealing that the company's claim to eliminate all such content was false. Knowledge discovery

may also take the form of long-term, embedded investigation. Pan (P16) describes the *"digital ethnography"* method that takes *"some tools from journalism and from forensic analysis"*: they *"enter into the communities, identify who the influencers are, and then we identify the type of techniques they use, and the type of strategies they use long term."* This type of investigation may occur over a period of six months or more.

Analysis serves multiple purposes. It helps mitigators assess the potential for harm or evaluate the threat severity of an incident to ultimately prioritize their efforts on higher-risk ones. For instance, data enrichment can guide mitigation teams to decide which incidents are potentially more harmful by identifying content from authors with a history of high impression volume per post [Rosa-P18]. The augmentation process can also lead directly to measurement of the impact of interventions. [Rosa-P18] explains the value of associating impressions with content:

> *"the top line number that we're trying to bring down in each domain setting, domains like Covid misinformation, spam, hate speech, electoral misinformation, in each of those categories... we're trying to estimate and reduce the number of impressions on that content. So actually it's not even that we're optimizing for the least amount of content possible, it's more like we're trying to have the least views of that content." —Rosa (P18)*

Knowledge discovery helps expand existing databases of known disinformers, known narratives and attacker behaviors, which in turn supports detection processes based on known indicators. Overall, analysis supports ongoing research to *"gain scientific understanding, ... look at larger patterns, [and] understand what generalizes to get a sense of how these things work, especially if we're going to think about solutions to mitigating mis- and disinformation"* [Hea-P8].

3) Mitigation: Mitigation takes many forms and is largely determined by the role or the organization of the participant. For example, the most common rapid mitigation response among participants is to report accounts and content for removal (n = 16). For trust and safety teams, longer-term responses may involve updating platform policies in response to emerging threat patterns (n = 6). Journalists and fact-checkers publish fact-checks as a rapid response, and they also perform longer-term investigative reporting to reveal disinformers and communicate findings from case studies (n = 9). Advocacy groups may advise clients on future public-relations (n = 2) or promote regulatory changes (n = 6).

It is important to note that participants also emphasized the importance of *doing nothing*. In some cases, mitigators wait and continue to monitor an emerging incident, to avoid inadvertently spreading or amplifying it themselves, where it might otherwise simply die down on its own (n = 3). Participants working in platform trust and safety or social network analysis also note that they wait temporarily for small or new incidents to develop further before intervening (n = 2). Gada (P7) explains this process:

> *"we talk a lot about ... the "tipping point," which is trying to understand, just because you can find a rumor it doesn't necessarily mean you should take action on it; so we have a set of metrics about, has it jumped platforms, how many shares has it got versus comments, is there an influencer that's been involved, what's the length of time that this has been circulating... we use those metrics to make a decision when we're talking to other partners about whether or not they should take action." —Gada (P7)*

In (Sec. IV-C), we suggest a threat characterization model which captures analytical factors that contribute to a threat's severity and structures them into a guiding framework. A systematic characterization of disinformation can facilitate automation, streamlining the detection, analysis, and mitigation functions currently performed by participants. A threat characterization model is also relevant for participants who currently rely on a less data-driven process of assessment and prioritization: for example, Ines (P9) describes their approach to predicting the spread of a given rumor as *"something that I do without thinking about it,"* and other participants who describe a more client- or funder-driven process for selecting which events to focus on (n = 6). Our threat framework offers a system which can be adopted at all stages of the disinformation incident life-cycle.

C. Threat Characterization

Developing an understanding of how disinformation actors operate is central to the effective mitigation of the associated risks. With this goal in mind, we characterize the threats of the disinformation landscape based on the hands-on experience of the experts and drawing inspiration from threat modeling practices within the security community [101], [103], [108].

In our framework, disinformation events or campaigns are characterized by the following four elements:

1) *Threat Actor*: Who creates, spreads or amplifies disinformation?
2) *Attack Patterns*: How do the actors effectively disinform?
3) *Attack Channels*: On which platforms and media do the actors disinform?
4) *Target Audience*: Who are the targets of the actors' attacks?

1) Threat Actor: A *threat actor* may be an individual, a group, or an organization that uses its resources to execute attacks and run campaigns on a target audience.

a) Sponsor and Agents: Threat actors broadly encompass different types of entities: sponsors and agents. *Sponsors* are individuals, groups and entities who are the source of a campaign and choose a narrative to be pushed. In their work, Omar (P15) makes reference to *"the 'ultimate sponsor,' ... the party who ordered that disinformation campaign to be spread." Agents* are actors who spread the elements of a campaign. Within this group, we distinguish *witting agents* and *unwitting agents*. *Witting agents* are informed actors who are aware of the presence of the disinformation campaign and intentionally participate in spreading and amplifying the narrative. *Unwitting agents*, on the other hand, are actors who are naive to the campaign and are unaware of their contribution to its goal [96].

b) Affiliation: Affiliation is another informative property which can be assigned to threat actors, as it is often correlated with other properties such as resources and capabilities. We define five categories which stand out in our findings: state, political, corporate, ideological and individual.

state State sponsored or affiliated actors often have motives aligned with national security, political, or commercial interests of the country of their origin. Multiple participants (n = 12) regularly observe these actors to be the front and center

of modern information operations. State involvement generally implicates complex political dynamics which are essential for mitigators to be aware of. Ehan (P5) also emphasized the need to avoid the *"othering"* of state actors: although some state actors are encountered more frequently, investigators cannot assume that certain states are never the threat.

political Actors with direct or indirect affiliations with domestic political parties are increasingly often identified (n = 9) behind disinformation campaigns, typically with the intent to expand their political influence and make electoral gains. Kai (P11) pointed out that disinformation is no longer limited to *"fringe groups discussing wacky theories,"* and has now entered the political mainstream, where parties are *"seeing the value in legitimizing misinformation"* to cause doubt and *"gain political capital out of it."* Hea (P8) explains that domestic activity constituted most of what they observed in the 2020 US elections, with *"well-known people repeatedly sharing false and misleading narratives that aligned with their political aims."*

corporate Multiple participants (n = 5) have observed an increase in information operations attributed to corporate actors, who are primarily motivated by economic interests and brand image. In the last two years of the Covid-19 pandemic, multiple operations have been run by various parties for *"the promotion of competing vaccines"* resulting in unfair market advantage to the perpetrators [Omar-P15]. Sam (P19) mentions seeing *"incorporated companies, LLC"* as actors behind disinformation campaigns.

ideological Activists aligned with ideologies, including conspiracy theories, actively rely on disinformation campaigns (n = 7) to promote and spread their agenda; this can result in serious danger to public health and safety in the process. They pose a particular challenge for mitigators as their commitment to their cause makes them especially persistent and effective at pushing narratives on their target audiences. Examples of such actors include anti-vax activists who strategically spread disinformation around the Moderna Covid-19 vaccine in Japan [Ines-P9], and QAnon believers who have pushed out campaigns inciting violence [Rosa-P18].

individual Actors can be unaffiliated and act in their individual capacities to pursue personal interests (n = 8). At the onset of the Covid-19 outbreak, before platforms had developed policies around the topic, Lak (P12) noticed individuals on their platform with the sole motivation of making *"a quick buck off of some really shitty [Covid] ads, that people are gonna click on."* Tay (P20) recalled that *"sometimes we see people [spread disinformation] just for their amusement."*

c) Motives: While the actors named by our participants have distinct affiliations, it is rare that only one motivation is involved in a campaign. As described by Tay (P20), many of the homegrown disinformation campaigns have *"a mix of political, financial, and personal promotion motivations . . . it's not as frequent to see one exclusive motivation behind a political campaign because they're profitable in many different ways."* Pan (P16) has observed financially-motivated actors operating in ideological communities such as anti-vax communities. They recall how *"a constellation of different communities"* within the larger anti-vax narrative included professionals such as lawyers, journalists, or politicians, motivated by *"an economic goal rather than ideological"*: selling products to credulous community members.

d) Resources, Capabilities & Sophistication: Threat actors vary in their access to resources and in their capabilities, which directly impacts the scale, turn over, and effectiveness of their operations.

A primary type of resource is financial: the financial resources available to threat actors strongly determine which attack patterns are available to them, and in general, more money allows for greater attack sophistication. Access to financial resources allows actors to build and execute build campaigns more quickly, by *"purchasing growth, whether that's advertisement or purchasing more followers or taking over accounts, whether that's renting them out or hacking them"* [Tay-P20]. Another resource for threat actors is their level of access to the distribution channels used to reach their target audiences. State actors may have control over media and news organizations, and as Marge (P13) explains: *"it becomes really tricky when . . . a reliable [media] source is operated completely by a government."*

Notably, our interviewees indicated human capital as a less-obvious resource which cannot be underestimated. Human capital also contributes to the strength of an attack, especially one which includes individual witting agents. While it is possible for well-entrenched actors to purchase organic behavior, an attack becomes far more robust when the people are committed to the cause. People who are strongly motivated by ideology may also build networks of like-minded actors which are particularly robust: Quin (P17) explains that *"if a network of far-right groups is removed one day, they are capable of creating new pages and new groups with hundreds and thousands of followers on the other day."*

Another important property of threat actors is their sophistication. Finn (P6) states it simply: *"if [actors] are able to develop sophisticated strategy, they are going to have a bigger impact."* Actors vary widely in their degrees of expertise and sophistication levels, and an actor's degree of sophistication may also evolve over time: *"we've seen [state actors'] tactics grow more and more sophisticated as a way to adapt to the mitigation measures that both platforms and also government agencies have put in place"* [Tay-P20]. Sophisticated actors develop resilience against mitigation by investing in *"diverse infrastructure that they can [use] to their benefit if they get shut off from one account [or] from one platform; they can still . . . keep going"* [Omar-P15]. Campaigns may become even more complex when multiple actors with varying levels of sophistication work together:

> *"We think about [sophistication] as a hierarchical problem . . . at the lowest tier . . . it's simple trolls or bots that work at a very large scale, but just spam the same message over and over. . . But, they will not work alone, they will work with more sophisticated actors who prime the target audience for that message, seed stories, can even infiltrate populations and become influencers in them and make them much more susceptible to the large scale messaging the less sophisticated actors undertake." —Babu (P2)*

2) Attack Patterns: Based on the campaigns and mitigation experiences described by our participants, we present 15 *attack tactics* of varying sophistication, from large-scale spamming with bots and cyborgs to generating realistic profiles and content with deep fakes. Some of these tactics are primarily *offensive* in nature, such as automatically generating opposition

rhetoric with the help of trolls. Some are primarily *deceptive*, such as generating realistic but fake pseudoentities. Others are primarily *evasive*, such as those that evade attribution. We group these tactics into six *attack patterns*. Table III summarizes the patterns, tactics and the number of participants who discussed them.

Pattern 1: Flood. This attack pattern aims to push a certain narrative by spamming a wide audience through the use of as much automation as possible. It includes the following tactics:

flood::bots Bots are autonomous programs that can run social media accounts to spread content without human involvement. Botnets are networks of bots that can interact with each other and coordinate posts with little or no attempt at persona development [89]. While some participants assumed varying degrees of automation in their description of bots depending on their technical background, many participants (n = 9) discussed the use of bots or botnets during recent events such as Brexit [Gada-P7], the 2016 US elections [Omar-P15], and the Venezuela elections [Ehan-P5]. Lak (P12) notes that bot detection is relatively easy for platforms as they have the *"technical data and infrastructure in place to capture and detect that sort of behavior"*; Ehan (P5) also considers botnet campaigns *"super easy to find."* Despite this, their modern usage in combination with other tactics can add complexity to a campaign which keeps them relevant for mitigation.

flood::cyborgs A cyborg is either a human-assisted bot or a bot-assisted human, inheriting characteristics from both [17], [20]. They initially produce automated responses before a human periodically takes over to produce more complex responses to user interactions: Rosa (P18) describes a cyborg as *"like a bot, but then if someone responds to them, a person will take over,"* and notes the increasing presence of these hybrid entities on their platform.

flood::copypasta Copypastas are text copied and pasted across the internet by individuals, usually at the same time. Different from something that is shared, copypasta can seem original without close examination [9]. Gada (P7) notes that copypasta was one vector of disinformation on the polio vaccine which spread on closed platforms such as WhatsApp.

Pattern 2: Drown. This attack pattern aims to hinder a group's ability to reach common ground by pushing inflammatory or incendiary content at all sides of a public debate, in order to drown out a specific view or create an environment more open to a particular message.

drown::trolls Trolls quarrel or upset users to distract and sow discord by posting inflammatory and digressive messages. The tactic takes a divide and conquer strategy, pitting the target group members against each other around heated topics [8], [61], [125]. Troll farms are organized online groups of agitators who identify divisions in other countries or groups, then insert themselves into those debates with the aim of inflaming. Multiple participants (n = 4) described the use of this tactic by Russian affiliated actors, such as the Internet Research Agency (IRA) [21], around heated topics in the US like the black lives matter movement [Hea-P8], gun control, and the vaccine mandates [Gada-P7]. Hea (P8) explains that IRA *"troll accounts were active on both sides ... of US political discourse ... trying to both infiltrate those different communities that were having conversations about Black Lives Matter and then shape those [conversations] towards their goals, rather than the goals of those communities."* While this tactic may be used by political actors, Tay (P20) also observes actors who troll just for their own fun and amusement: *"they do look to impact the conversation, [but] they don't necessarily always look to impact the conversation in a way that builds political capital for them personally."*

Type of Pattern	Pattern::Tactic	Frequency
Offensive Patterns	flood::bots	9
	flood::cyborgs	1
	flood::copypasta	1
	drown::trolls	4
	drown::hijacking	2
Deceptive Patterns	counterfeit::pseudoentities	10
	counterfeit::astroturfing	3
	counterfeit::pseudocontent	4
	infiltrate::seed-invite-amplify	3
	infiltrate::mainstream	11
Evasive Patterns	evade-detection::gaming heuristics	3
	evade-detection::ML poisoning attack	1
	evade-detection::crowdsource	2
	evade-attribution::proxy companies	1
	evade-attribution::dark PR firms	1

TABLE III: Attack patterns with tactics and the number of participants who mention them. With an issue as complex and diverse as disinformation, we find value in reporting tactics even if they were mentioned once.

drown::hijacking The purpose of this tactic is to hijack a trend or cause in order to promote one's own narrative and agenda (n = 2). *Hashjacking*, the use of someone else's hashtag to promote one's own agenda, is known to polarize communities on Twitter [19]. Rosa (P18) explains that certain regimes manufacture consensus for their actions within social media platforms by *"hijacking any attempts by alternative voices and drowning them out essentially on social platforms."* To explain the drowning of a specific view, Omar (P15) used the example of *"an oil company in Brazil or Peru [that tries] to put down or stifle an indigenous protest against drilling using social media."* Rosa (P18) mentioned the use of this tactic by corporate actors to *"drown out a negative trend."*

Pattern 3: Counterfeit. This attack pattern consists of campaign tactics which involve creating fake identities or organizations, falsely simulating popular support, and injecting content that appears deceptively real, with the goal of enhancing the credibility of the disinformation. Multiple participants (n = 5) emphasized the importance of source credibility in effectively deceiving a target demographic and in evading detection and mitigation measures.

counterfeit::pseudoentities Unlike automated flooding tactics such as bots and cyborgs, this tactic invests significantly more effort and resources to create realistic fake entities. For example, *sock puppets* are multiple online identities controlled by a single party, often for purposes of deception, to fulfill goals such as supporting a cause, changing policies, manipulating online opinions, or circumventing restrictions (n = 7).

Participants have also encountered the use of *off-platform* resources to grant legitimacy to these fake identities (n = 3):

Ehan (P5) describes a network of fake personas posing as Americans and deriving credibility through a fake website in a Russian-backed disinformation campaign, and Quin (P17) explains how a Russian-backed campaign created entertainment websites and Facebook accounts in the Georgian language.

Our participants also describe how fake personas with *information roles*, such as journalists and think tank members appear more credible (n = 4). As Finn (P6) explains, *"if you want people to read you, it's easier to impersonate the media or journalists ... than anything else, because people are looking at these kinds of actors to collect information."* These personas do not need to belong to real information organizations: Quin (P17) observes an increase in the creation of fake websites that look like news sites but have a specific political agenda. Omar (P15) explains that *"Russia sets up fake think tanks in different countries like Serbia or even some countries in Africa"* to interfere with Ukrainian and African elections. Finn (P6) notes how one campaign created a fake online magazine issued by the European Parliament.

counterfeit::astroturfing Astroturfing as a tactic aims to create an illusion of a genuine grass-roots support or opposition to a group or a policy, through centrally-coordinated witting agents that appear to be independent and ordinary citizens (n = 3) [54]. The identity of the sponsor is intentionally distanced from the mobilization effort. Sam (P19) mentioned repeated incidents of *"corporate astroturfing"* by corporate actors, such as tobacco, energy and insurance companies. State actors also use such tactics to manufacture consensus, making it seem *"like everyone around you is in support of whatever government action [has been taken]"* [Rosa-P18].

Astroturfing attacks may co-opt platforms' popularity mechanisms, such as trending topics, where chosen keywords or topics are artificially promoted by coordinated and inauthentic activity to appear popular [22]. Rosa (P18) has encountered the tactic in use by well-resourced threat actors who purchase trending topics to emulate wide scale support for their cause.

counterfeit::pseudocontent This tactic creates deceptively realistic fake content by manual or automated methods [109], [129]. Our participants observe a large variance in sophistication employed to create fake content: from simple-yet-effective *click baits* that attract users to follow links to articles containing misinformation (n = 2), to *cheap fakes* (n = 2) generated with unsophisticated technology such as reusing stock images or existing profile pictures, to the use of *deep fakes* (n = 3), in which a person in an existing image or video is replaced with someone else's likeness to create hyper-realistic content using deep learning models [117]. Highlighting the deceptive capabilities of AI-generated content, Noel (P14) commented that *"one out of three deepfakes is not properly identified."* Contrary to their expectation that deep fakes would appear as a standalone category in the 2020 US elections, Lak (P12)'s investigations revealed that deep fakes did not appear *"in isolation and were very much partnered with a misinfo or disinfo narrative."*

Pattern 4: Infiltrate. Unlike the counterfeiting attack pattern, which relies on fake personas, fake entities, and manufactured coordination, this pattern relies on influencing normal users to themselves create and spread disinformation.

infiltrate::seed-invite-amplify In this tactic, a campaign invites normal users to engage with a *seed* misinformation incident (n = 3). Ehan (P5) explains that Russian-backed campaigns would often *"actively search for engagement, ... [by] telling readers [to] come to see what they are posting on the website and give their opinion, interact ... spread stuff and so on."* Hea (P8) describes a case seen during the 2020 US elections:

> *"We could see political leaders and media leads kind of pushing [seed] this frame that there was going to be voter fraud ..., and then we can see people on the ground or everyday people pick up these frames of expecting voter fraud, and then they would misinterpret what they were seeing in the world and create [invite] their own false and misleading narratives from their own experiences. So it wasn't explicitly coordinated, it has ... organic components. And then influencers would opportunistically retweet [amplify]." —Hea (P8)*

infiltrate::mainstream In this tactic, actors involve media, politicians, celebrities, influencers, and bloggers in the target audience such that the message appears mainstream (n = 11). Some actors achieve the mainstreaming of their message by *"becoming influencers in the [target population] and making them much more susceptible to the large scale messaging"* [Babu-P2]. Tay (P20) also talked about the appropriation of existing influencers, who then *"spread the false information, or not even necessarily false information, but sometimes just decontextualized information on behalf of an actor."* Actors with adequate financial resources may even involve real, unwitting journalists in constructing their fake media sources [Ehan-P5]. Tay (P20) notes that manipulators deliberately involve mainstream media because it *"lends credibility to the false information in a way that even most popular online influencers cannot."* They describe a case demonstrating the power of mainstream media to amplify disinformation:

> *"[It] started out as one single blog post in a small county that was then picked up by Republican politicians within that county, that then trickled up through more mainstream legitimized media like Newsmax, OAN and Fox News up to the President, and then was again re-disseminated through more traditional media throughout the US voting public." —Tay (P20)*

Pattern 5: Evade Detection. This evasive pattern consists of tactics that enable a campaign to evade detection long enough to achieve its goals.

evade-detection::gaming heuristics Detection algorithms often rely on simple heuristics and policies (n = 3). Threat actors aim to *"circumvent algorithmic protections deliberately and thoughtfully"* [Lak-P12]. Lak (P12) describes this tactic as a *"cat and mouse game"*: if actors cannot say the word "Covid-19" on a YouTube channel for fear of being instantly demonetized, they can replace it with a code word which their audience will recognize, but an algorithm will not. Rosa (P18) explains how one can *"build [a flagged word] with emojis or build it with some kind of character replacement,"* to avoid getting caught by simple keyword filters.

evade-detection::ML poisoning attack Machine learning models are increasingly used by platforms to automatically filter misinformation. A poisoning attack occurs when the adversary injects specifically engineered data into a model's training dataset which causes the model to learn a manipulated mapping. Threat actors can use such attacks to modify classification output and produce their desired false result [55]. Rosa (P18) explained that this is a *"classic risk involved with using ML tools"*: threat actors can effectively

inject engineered data by performing behavior which causes the model to *"learn something based on artificial or adversarial actions and then just kind of go nuts."* This is a "backdoor" which the attacker can use for instance to cause a model to classify a post as factual if it contains a certain word [14].

evade-detection::crowdsource Similar to counterfeiting tactics, crowdsourcing relies on embedding realistic entities and behavior to not only avoid detection but also to avoid breaching platform policies that ban synthetic accounts such as bots (n = 2). As Omar (P15) explains, actors

> *"will circumvent moderation efforts that tackle coordinated inauthentic behavior with authentic behavior; they will hire and they will build seemingly authentic entities ... they will pay actual people to help them spread disinformation, because they know it will be very hard for coordinated inauthentic behavior policies to actually run them on the fly." —Omar (P15)*

Omar (P15) also notes that actors can build extensive offline, off-platform assets: they *"start offline with real people and [then] go online to different platforms . . . [they] pay people in India or in the Philippines $1 a day to promote something."*

Pattern 6: Evade Attribution. This evasive pattern aims to hide the identity of the attack sponsors and make attribution more challenging.

evade-attribution::proxy companies In this tactic, an actor pays one or more proxy companies to front their campaign: Omar (P15) explains, *"it's not building a bot farm in St Petersburg, it's hiring a company that hires another company that hires another company to do it on behalf of a state actor or a corporation."*

evade-attribution::dark PR firms In this tactic, an actor makes use of public relations (PR) firms specialized in providing existing infrastructure as a service to clients looking for quicker and cheaper setup. Omar (P15) gives the example of the Argentinean presidential elections where a *"Spanish-speaking PR firm that [had] worked for a customer in Spain repurposed accounts for an Argentinean audience."* They explain that this type of off-platform resource is generally used by *"political parties and not [by] a tier one threat actor like Russia and China, Cuba, North Korea, Iran."*

3) Attack Channels: Our participants describe four primary channels where they investigate or observe disinformation activity, with examples listed in Table IV.

social media platforms All participants describe disinformation activity on social media platforms. Babu (P2) notes that social media is a *"very powerful place"* where a *"small group of actors"* are able to *"target a lot of different populations very quickly."* Some participants (n = 3) discuss that each platform has a *"different presence in each region"* [Udo-P21]. This in turn determines an actor's choice of platform in the region. Quin (P17) refers to Facebook as *"the main war theater"* in Georgia: *"the majority of Georgians are present on Facebook and they receive their daily information from the platform . . . that's why these actors are present on Facebook and they try to invest in it a lot."* They contrast this with activity on Twitter, which is less popular in Georgia and thus a lower priority for actors.

messaging platforms Almost half of our participants (n = 10) investigate disinformation campaigns in closed, semi-closed, anonymous, or semi-anonymous online messaging platforms. Omar (P15) describes how political parties in Latin America *"launch targeted disinformation campaigns [on] WhatsApp [or] Telegram by obtaining phone numbers of voters."* Gada (P7) points out that while investigators are mostly focused on social media platforms, *"the biggest problem is health and science misinformation on closed messaging apps."* Tay (P20) and Vera (P22) also find in their experience that coordinated campaigns start on this channel.

social media platforms		messaging platforms		news media	
Example	#	Example	#	Example	#
Facebook	18	Telegram	6	Fox News	3
Twitter	18	WhatsApp	6	OAN, Reuters, CNN,	1
YouTube	11	Discord, Signal	1	Russia Today, CNBC,	
TikTok	4			Newsmax, Bloomberg	
Instagram	3				
Wikipedia	2				
LinkedIn, Parler, 4chan, Snapchat, Quora, Gab, VK	1				

TABLE IV: Example platforms and media in three of the main attack channels as listed by participants.

news media Half of our participants (n = 11) discuss the role of online and offline mainstream media (TV and print news companies) in legitimizing disinformation. Tay (P20) notes that *"mainstream media has become such a target of false and misleading campaigns, because the manipulators generally know that if the media says something it becomes more important and more credible than if it just travels throughout the web."* Chan (P3) also emphasizes that TV is a *"big issue"* and that *"a lot of disinformation which has an absurd impact in a country like Italy passes through television."* Similarly, Kai (P11) discusses the *"damage that outlets like Fox or Russia Today are doing to many international discussions around climate change."*

websites Several participants (n = 7) mention the use of websites, often promoted on social media and messaging platforms, as a channel to spread disinformation. Pan (P16), who studies communities on Facebook and Telegram, explains how actors aimed *"to push the people onto websites [where] they were constantly asking for donations, selling masks, products, and, more dangerously, . . . selling at-home therapies."* Finn (P6) observes that *"[disinformation] often starts with websites because actors need to have credibility . . . it's easier when you have a website."*

Modern disinformation operations often make use of multiple channels simultaneously to achieve their goals. Several participants (n = 8) highlight this cross-platform nature: *"it used to be that we could study a campaign on just one platform, but increasingly, we need to study a campaign on Twitter, Facebook, Telegram and other smaller platforms, and mainstream media or online mainstream media"* [Babu-P2]. Vera (P22) observes the *"cross pollination of mis- and disinformation"* from *"fringe platforms or the dark web or closed messaging networks"* to mainstream ones. Ehan (P5) talks about a disinformation group that was exposed on one platform, but were later found *"active on Gab and Parler, trying to find new ways to build a community where they're going to spread their content."* Tay (P20) makes a similar observation: *"deplatformings have pushed some of the malicious actors*

to alternate platforms, whether that's establishing their own platforms or using existing platforms to rebuild their audiences and continue spreading false information, to various degrees of success."

4) Target Audience: Disinformation campaigns seek to cause harm by influencing recipients of disinformation: their target audience. The choice of audience ("who?") can enhance the effectiveness of a campaign, and the chosen audience in turn determines other strategic choices ("how?") such as the selection of attack patterns and channels. Harmfulness of a campaign does not depend solely on the technical capabilities and resources of the threat actor: *"something might be harmful because it is particularly damaging to a vulnerable population"* [Babu-P2]. While any audience may be targeted by disinformation, threat actors often develop strategies based on several key traits which contribute to the susceptibility of an audience.

demographic Several participants (n = 5) mention that demographic characteristics play a role in the choice of a target audience. These characteristics include, but are not limited to, age, gender, religion, nationality, ethnicity, or professional status. Participants encounter targeting of groups based on religion (e.g., Muslims in India [Ehan-P5]), sexual orientation (e.g., LGBTQ in Georgia [Quin-P17]), age (e.g., youth during protests [Omar-P15]), gender (e.g., women in politics or holding public office [Chan-P3]), and ethnicity (e.g., Cuban Americans [Gada-P7]). Threat actors can maximize the impact of a campaign by choosing their attack channel based on the demographics of the target audience, as in Russia's use of TikTok to target youth for involvement in protests [Omar-P15].

digital literacy The digital literacy of the target audience can determine their susceptibility to disinformation narratives (n = 3). Babu (P2) explains, *"a public that already has a high level of sophistication versus a public that does not have a lot of exposure or understanding of disinformation ... can certainly factor into how harmful or how impactful that campaign might be."* The target audience's *"information resources and technology literacy"* [Babu-P2] inform the toolkit of attack patterns deployed by threat actors. Gada (P7) uses what they term the *"information diet"* of a community as an indicator of its vulnerability to misinformation, naming properties such as high usage of closed messaging apps and low levels of news consumption as markers of susceptibility.

fact-checking capacity The quality of fact-checking resources available to a target audience also impacts how susceptible the audience is to disinformation campaigns (n = 5). When determining the severity of threat for a particular audience, Babu (P2) asks, *"are there public agencies in the target population whose job it is to fact check or verify social media? If so, how effective are they?"*. Quin (P17) explains that *"one of the problems that Georgia faces is the lack of good investigative journalism which would work not only with open sources, but in the Bellingcat[4]-style investigation."* The language spoken by the target audience is also a factor in fact-checking capacity. Overall, fewer tools and resources are dedicated to less-common languages: given that resources are limited, fact checkers prioritize larger-scale languages and platform integrity teams prioritize larger markets. Quin (P17) captures this limitation: *"the tools we use are focusing on the most-spoken languages, like English, Russian, Chinese ... it is hard to use them when covering the less-spoken languages."*

[4]Bellingcat is a Netherlands-based investigative journalism website specialising in fact-checking and open-source intelligence.

V. Applying the Threat Model

A. Case Studies

Our proposed threat model provides a thorough and comparison-friendly articulation of disinformation threat scenarios. To demonstrate its applicability, we select six disinformation campaigns, uncovered within the last two years, that are publicly accessible as case study reports. For each of these examples, we map out the attributes of threat actors at play, the attack patterns they deployed, the attack channels they chose, and the audiences they targeted. Table V displays the results of the threat characterization. We provide more details on the application of our framework to these campaigns.

Example 1: Russia targets US Far Right through unwitting journalists.[5] Russian state-affiliated actors ran a fake news website to attract right-wing journalists to target American users with pro-Trump and anti-Biden messaging, and infiltrated far-right audiences on Gab and Parler to push the users toward both ends of the political spectrum with hyper-partisan content. Our threat characterization yields that the actors' patterns indicate state affiliation, and their tactics include commissioning journalists, hinting at their desire to mainstream their narratives. Their choice of Gab and Parler social media platforms takes advantage of these platforms' lack of content moderation, and their choice of the far-right as the target audience results from the susceptibility of this demographic to their narrative.

Example 2: Pro-India group discredits Pakistan in the EU.[6] A Geneva-based disinformation network, spread over 100 countries during its 15 years of operation, resurrected a dead professor, revived over 10 defunct UN-accredited NGOs, and manufactured over 750 fake media outlets to discredit Pakistan and influence decision makers at the UN and European Parliament. Characterizing the operation with our model highlights a state actor, its reliance on fake NGOs and think tanks (pseudoentities), on coordination with India's largest wire service ANI (mainstream), and on mobilization of Geneva-based students for demonstrations (crowdsource). Their successful execution of this campaign on a target audience with a sophisticated digital literacy and an established fact-checking capacity reveals the actors' advanced skills and capabilities.

Example 3: Constellation of anti-vaccine conspiracy theories take hold in West Africa.[7] A collection of domestic and foreign actors are spreading anti-vaccine narratives in West Africa, using content sourced from North American (QAnon) and European (French disinformation websites) conspiracy groups, with the goal of eroding trust in the institutions and disrupting vaccination efforts in the region. Applying

[5]https://www.reuters.com/article/us-usa-election-russia-disinformation-ex-idUSKBN26M5ND

[6]https://www.bbc.com/news/world-asia-india-55232432

[7]https://firstdraftnews.org/long-form-article/foreign-anti-vaccine-disinformation-reaches-west-africa/

Domain	Actors: Sponsors	Agents	Motive	Affiliation	Attack Patterns: Tactics	Specifics	Channels: e.g.	Targets: demographic	dl	fc
Democracy Ex. 1	IRA	Right-wing journalists	Rally pro-Trump support	state	drown::trolls counterfeit::pseudoentities counterfeit::pseudocontent infiltrate::mainstream infiltrate::seed-invite-amplify	Inflame racial tensions Far right organizations Fake personas/websites Deep fake photos Commission journalists Invite user-interactions	social media news	Far-right Americans		●
National Security Ex. 2	Pro-India Network	EU representatives	Undermine Pakistan's credibility	state	drown::hijacking counterfeit::pseudoentities infiltrate::mainstream evade-detection::crowdsource	Hijack minority issues 750+ outlets; 10+ NGOs Wire-service coordination Involve Geneva-based students	social media news web	UN & EU Parliament members	●	●
Public Health Ex. 3	QAnon	Local social media users	Disrupt vaxx efforts	ideological	flood::copypasta counterfeit::pseudocontent	Posts in quick succession News modification	social media messaging	West Africans		
Economy Ex. 4	Not found/ Insufficient evidence	Huawei executives	Anti-Belgian govt plan for Huawei	corporate	evade-attribution::proxy companies counterfeit::astroturfing infiltrate::mainstream flood::bots counterfeit::pseudoentities	Unattributable origin Mimic organic support Invite Huawei executives Amplify with bots GAN (AI) profile photos Create & amplify articles	social media web news	Western European audiences	●	●
Public Safety Ex. 5	VDARE Unz-Review	White nationalists	Advance racial stereotypes	ideological	flood::copypasta counterfeit::pseudocontent	Coordinated postings Divert to off-platform sites Systematic amplification using inauthentic accounts	social media web	White American audiences		●
Public Safety Ex. 6	Myanmar Military members	Pro-army socia media users	Support military-backed opp. party	political	counterfeit::astroturfing drown::hijacking flood::cyborgs counterfeit::pseudocontent	Intense activity bursts Downplay Rohingya genocide Fake accounts Fb Pages sharing news Impersonation of celebrities	social media news	Ruling political party		

TABLE V: Application of our threat characterization model to six disinformation campaigns. '●' indicates the existence of adequate digital literacy (dl) or fact-checking capacity (fc) in the target demographic.

our framework, we find that ideological actors are exploiting the historic vaccine hesitancy in the target audience, whose lack of digital literacy and poor fact-checking capacity makes them susceptible to tactics such as copypasta across social media platforms and messaging platforms.

Example 4: Inauthentic accounts target Belgian Government's plans to limit Chinese firms.[8] A cluster of inauthentic accounts attacked the Belgian government's plan to limit access of Chinese firms, notably Huawei, to its 5G network. Our threat characterization yields that the actors' patterns indicate corporate affiliation, and their tactics include astroturfing by mimicking support through articles and posts in various European languages, reaching mainstream audiences by inviting Huawei executives to interact with their online posts, and setting up bots supported by GAN-generated profile photos. They amplified their narrative among west European demographic on social media by sharing content from a combination of handpicked news and web sources.

Example 5: White nationalist group advances racial stereotypes by inorganically amplifying books and websites.[9] Anti-immigrant groups, VDARE and Unz Review, pushed their ideological agenda of attacking people of color among their target demographic of white Americans. One of their tactics included easier-to-detect copypasta postings of the same content in the same sequence within a time span of a few minutes. They also relied on coordinated amplification of pseudocontent hosted almost exclusively at three web pages.

Example 6: Myanmar military assets engage in PR and inflate support for opposition party before elections.[10] Through their social media agents, members of the Myanmar military sponsored a campaign that actively propagated pro-army and pro-opposition political narratives and targeted the ruling political party demographic. Through periods of intense posting, the campaign performed astroturfing to show wider support, hijacking alternate voices on Rohingya genocide by pushing the army's stance, impersonated celebrities and social media influencers to provide credibility to their pseudocontent. Since Facebook is the dominant form of social media in Myanmar, the campaign focused primarily on this platform, supplemented by some assets on Instagram.

B. *Utility and Anticipated Usage*

The systematic framework facilitated us—and is anticipated to facilitate mitigators—to better organize unstructured information about disinformation campaigns into a compact, structured form that is communicable to a diverse set of stakeholders and conducive to understanding and comparing different operations.

Toward standardized, efficient analysis: Whereas multiple experts in our study appreciated the need for a cybersecurity-inspired approach to analyzing disinformation campaigns, they

[8] https://public-assets.graphika.com/reports/graphika_report_fake_cluster_boosts_huawei.pdf

[9] https://public-assets.graphika.com/reports/graphika_report_vdare_takedown.pdf

[10] https://public-assets.graphika.com/reports/graphika_report_myanmar_military_network.pdf

identified the lack of in-house expertise as an obstacle to realizing this goal. Indeed, as highlighted in Table I, mitigators working at the forefront of disinformation campaigns have varying levels of expertise in threat modeling. Our proposed framework is well positioned to bridge such knowledge gaps and may be used by analysts to ensure comprehensive coverage of different aspects of campaigns by prompting them to look for each dimension of the taxonomy. The framework assists the non-security community in its treatment of disinformation threats, and it is also well-placed to facilitate follow up research in the security community on this important problem.

Toward an automated procedure: As pointed out by multiple experts in our study, one major obstacle to the effective mitigation of threats is resource constraint: teams have too much content to monitor, and the lack of bandwidth to respond quickly means that a harmful narrative can go viral faster than teams can intervene, resulting in more extensive damage. Following threat characterization, it is standard cybersecurity practice to quantify the severity of threats as a means of triage [65]; such a numerical scoring system could be used to rank disinformation campaigns and guide the work of mitigators by helping them prioritize incidents by severity. Automation will be essential to implementing our model at scale: to develop effective threat assessment and triage systems built on top of our model, it will be crucial to test the model on a large set of disinformation campaigns, which will in turn require semi-automated processes to fill the model with concrete campaigns, in addition to detecting attack strategies. While the development and feasibility evaluation of automated detection techniques is outside the scope of this paper, we offer several suggestions of framework components with potential for automation, and related work on relevant methods, summarized in Table VI. Our investigation reveals actively researched directions toward automation of most of the components of the framework. These techniques can be leveraged to semi-automate the application of the framework for concrete campaigns, possibly in real-time.

Toward tackling cross-platform campaigns: Our application of the framework shows that disinformation campaigns are increasingly conducted in a cross-platform setting. This is in line with recent research [4], [33], [43], [113] showcasing the magnitude of this phenomenon. Our framework actively encourages the analysts to take a broader view in their mitigation effort by capturing different channels involved in the modern cross platform operations.

Toward capturing blended disinformation tactics: Our analysis of the case studies reveals that many of the campaign tactics are rarely utilized in isolation but rather in combination to achieve the desired goals of the operation. Blended disinformation campaigns use a combination of multiple attack patterns and tactic capabilities to achieve their ultimate goal. Such blended activity draws parallels to malware operations in practice, where a combination of malware capabilities are leveraged to perform complex attacks, spreading rapidly and infecting multiple endpoints quickly. Similar to malware behavior classification systems [82], a framework to capture disinformation is bound to have overlap in some of the categories due to the various goals the underlying tactics attend to. Our proposed framework is intentionally designed to be flexible to capture the complex patterns at play.

Component	Subcomponent	Approaches
Actors	Agents	[1], [34], [93], [116]
	Affiliation	[92], [94]
Offensive Patterns	bots	[18], [58], [67], [72]
	cyborgs	[73], [78], [88]
	copypasta	[100]
	trolls	[25], [60], [83], [104]
	hijacking	[49], [69], [106]
Deceptive Patterns	pseudoentities	[63], [107], [120]
	astroturfing	[42], [76]
	pseudocontent	[24], [46], [111], [112]
	seed-invite-amplify	[2], [116]
	mainstream	[34], [39], [92]
Evasive Patterns	gaming heuristics	[45]
	ML poisoning attack	[48], [75]
Channels	social media	[93], [98], [119]
	web	[15], [45]
	news	[7], [129]
	messaging	[28]
Target	demographic	[13], [32]

TABLE VI: Towards Automation: a selection of framework components for which technical approaches with automation potential are actively researched and developed. Determination of other components requires active human-in-the-loop involvement or manual off-platform investigations.

VI. Open Research Questions

The design and application of our framework indicates further directions of research which build upon it. Our work may provide a starting point for developing solutions to open questions at each of the identified stages in mitigators' work:

Detection: Forecasting when risk becomes threat. Determining when suspicious activity develops into an actual threat is not straightforward, and while signals like reach or virality of content can provide initial leads, according to Dany (P4), *"that's not really how risk turns into actual threats . . . [for example] the threat to life to an executive or a senior government official, it might not have the greatest reach in the world, but it's a very significant threat."* Further research can explore ways of combining raw signals with a framework such as ours which focuses on higher-level campaign concepts.

Analysis: Quantifying "impact." Many participants (n = 9) expressed the desire for a more structured process of measuring the "impact" of a disinformation campaign. Chan (P3) shares, *"One of the great issues that we have is to assess the impact of a single piece of disinformation"*; Hea (P8) expresses that *"impact is the million dollar question; it's actually really hard to measure the impact of one misinformation campaign."* Using our framework to precisely identify the elements and patterns of campaigns lays groundwork for assigning scores to individual events and composing them to assess a campaign overall. Similarly, as Tay (P20) describes, determining a population's vulnerability is *"really tricky . . . there should be more research in the area of formally quantifying it."* Properties like

demographic, digital literacy, and fact-checking capacity can be useful proxies in assessing potential audience vulnerability and moving toward more formal quantification.

Analysis: Exposing the ultimate sponsor. An important research question in *knowledge discovery* is understanding the sponsors behind a campaign, gaining insights into their motives and capabilities in order to better understand the threat landscape. Attribution is a hard problem, and sometimes *"the only way [it] can be done is to prove a financial link between those authentic threat actors"* [Omar-P15]. However, our framework can help mitigators to classify actors and specify their capabilities, which can assist with identifying when multiple campaigns may share a common sponsor, or tracking patterns and change over time in the activities of different actor types.

Mitigation: Informing platform response. The current variation in how platforms respond to disinformation activities is understudied and the understanding could guide the development of a universal, platform-agnostic scoring framework: *"the same campaign will be on five different platforms and they will take five different sets of actions against it ... I think it would be very, very important ... for the practitioners in the field to understand how the platforms are responding to different campaigns"* [Babu-P2]. Udo (P21) emphasizes the need for platforms to update their policies *"in real time"* in response to constantly evolving tactics and trends. Future work could connect the properties of a campaign with platform response and outcome, for instance comparing similar campaigns with different mitigations and outcomes, or differences in platform response in cases of cross-platform campaigns. This can advance understanding of which mitigation efforts are more successful in different cases.

VII. Conclusion

Based on interviews with disinformation experts, we present deep insights into the day-to-day functions of their fight against disinformation. We characterize the disinformation threat across domains by mapping out potential threat actors, their motives and capabilities, their observed patterns of attack, the attack channels they use, and the audiences they target. Our disinformation threat framework is a crucial step toward comprehensively understanding the attacker side, which is a necessary foundation for developing effective tools, methodologies, and countermeasures against disinformation.

Acknowledgment

This work was supported by NYUAD's Centers for Cyber Security (CCSAD) and Interacting Urban Networks (CITIES funded by Tamkeen under the Research Institute Award CG001), ASPIRE AARE-2020-307 and supported by CHIST-ERA within the CIMPLE project (CHIST-ERA-19-XAI-003).

References

[1] M. Alassad, M. N. Hussain, and N. Agarwal, "Finding fake news key spreaders in complex social networks by using bi-level decomposition optimization method," in *International Conference on Modelling and Simulation of Social-Behavioural Phenomena in Creative Societies*. Springer, 2019, pp. 41–54.

[2] M. Aldwairi and A. Alwahedi, "Detecting fake news in social media networks," *Procedia Computer Science*, vol. 141, pp. 215–222, 2018.

[3] S. Ali, M. H. Saeed, E. Aldreabi, J. Blackburn, E. D. Cristofaro, S. Zannettou, and G. Stringhini, "Understanding the effect of deplatforming on social networks," in *Web Science Conference*. ACM, 2021, pp. 187–195.

[4] M. Aliapoulios, A. Papasavva, C. Ballard, E. De Cristofaro, G. Stringhini, S. Zannettou, and J. Blackburn, "The gospel according to q: Understanding the qanon conspiracy from the perspective of canonical information," *arXiv preprint arXiv:2101.08750*, 2021.

[5] H. Allcott, M. Gentzkow, and C. Yu, "Trends in the diffusion of misinformation on social media," *Research & Politics*, vol. 6, no. 2, 2019.

[6] A. Badawy, K. Lerman, and E. Ferrara, "Who falls for online political manipulation?" in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 162–168.

[7] R. Baly, G. Karadzhov, D. Alexandrov, J. Glass, and P. Nakov, "Predicting factuality of reporting and bias of news media sources," *arXiv preprint arXiv:1810.01765*, 2018.

[8] S. Barsotti, "Weaponizing social media: Heinz experts on troll farms and fake news," *https://www.heinz .cmu.edu/media/2018/October/troll-farms-and-fake-news-social-media-weaponization*, 2018, accessed: 2022-02-01.

[9] BBC News, "The world of misinformation and fake news is full of confusing vocabulary - beyond fake news," *https://www.bbc.co.uk/beyondfakenews/fakenewsdefinitions/*, accessed: 2022-02-01.

[10] S. Bradshaw, H. Bailey, and P. N. Howard, "Industrialized disinformation: 2020 global inventory of organized social media manipulation," *Computational Propaganda Research Report, University of Oxford*, 2020.

[11] T. Burki, "Vaccine misinformation and social media," *The Lancet Digital Health*, 2019.

[12] K. M. Caramancion, "An exploration of disinformation as a cybersecurity threat," in *2020 3rd International Conference on Information and Computer Technologies (ICICT)*. IEEE, 2020, pp. 440–444.

[13] N. Cesare, C. Grant, and E. O. Nsoesie, "Detection of user demographics on social media: A review of methods and recommendations for best practices," *arXiv preprint arXiv:1702.01807*, 2017.

[14] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[15] Z. Chen and J. Freire, "Proactive discovery of fake news domains from real-time social media feeds," in *Companion Proceedings of the Web Conference 2020*, 2020, pp. 584–592.

[16] R. Chesney and D. Citron, "Deepfakes and the new disinformation war: The coming age of post-truth geopolitics," *Foreign Aff.*, vol. 98, p. 147, 2019.

[17] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who is tweeting on twitter: human, bot, or cyborg?" in *Proceedings of the 26th annual computer security applications conference*, 2010, pp. 21–30.

[18] S. Cresci, "A decade of social bot detection," *Communications of the ACM*, vol. 63, no. 10, pp. 72–83, 2020.

[19] P. Darius and F. Stephany, "How the far-right polarises twitter:'hashjacking'as a disinformation strategy in times of covid-19," in *International Conference on Complex Networks and Their Applications*. Springer, 2021, pp. 100–111.

[20] R. Di Pietro, S. Raponi, M. Caprolu, and S. Cresci, "New dimensions of information warfare," in *New Dimensions of Information Warfare*. Springer, 2021, pp. 1–4.

[21] R. DiResta, K. Shaffer, B. Ruppel, D. Sullivan, R. Matney, R. Fox, J. Albright, and B. Johnson, "The tactics & tropes of the internet research agency," *https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1003&context=senatedocs*, 2019.

[22] T. Elmas, R. Overdorf, A. F. Özkalay, and K. Aberer, "Ephemeral astroturfing attacks: The case of fake twitter trends," in *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2021, pp. 403–422.

[23] EU Disinfo Lab, "Why disinformation is a cybersecurity threat," *https://www.disinfo.eu/advocacy/why-disinformation-is-a-cybersecurity-threat/*, 2021, accessed: 2022-02-01.

[24] T. Fagni, F. Falchi, M. Gambini, A. Martella, and M. Tesconi, "Tweepfake: About detecting deepfake tweets," *Plos one*, vol. 16, no. 5, p. e0251415, 2021.

[25] P. Fornacciari, M. Mordonini, A. Poggi, L. Sani, and M. Tomaiuolo, "A holistic system for troll detection on twitter," *Computers in Human Behavior*, vol. 89, pp. 258–268, 2018.

[26] C. François, "Actors, behaviors, content: A disinformation abc highlighting three vectors of viral deception to guide industry & regulatory responses," *Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression*, 2019.

[27] C. François, B. Nimmo, and C. S. Eib, "The IRA CopyPasta Campaign," *Graphika, okt*, 2019.

[28] J. Gaglani, Y. Gandhi, S. Gogate, and A. Halbe, "Unsupervised whatsapp fake news detection using semantic search," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2020, pp. 285–289.

[29] J. Galvao, "Covid19: the deadly threat of misinformation," *The Lancet Infectious Diseases*, 2020.

[30] K. Garimella and D. Eckles, "Images and Misinformation in Political Groups: Evidence from WhatsApp in India," May 2020, arXiv:2005.09784.

[31] C. Geeng, S. Yee, and F. Roesner, *Fake News on Facebook and Twitter: Investigating How People (Don't) Investigate*. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–14.

[32] S. Golder, R. Stevens, K. O'Connor, R. James, G. Gonzalez-Hernandez *et al.*, "Methods to establish race or ethnicity of twitter users: Scoping review," *Journal of Medical Internet Research*, vol. 24, no. 4, p. e35788, 2022.

[33] Y. Golovchenko, C. Buntain, G. Eady, M. A. Brown, and J. A. Tucker, "Cross-platform state propaganda: Russian trolls on twitter and youtube during the 2016 us presidential election," *The International Journal of Press/Politics*, vol. 25, no. 3, pp. 357–389, 2020.

[34] S. Guarino, N. Trino, A. Chessa, and G. Riotta, "Beyond fact-checking: Network analysis tools for monitoring disinformation in social media," in *International conference on complex networks and their applications*. Springer, 2019, pp. 436–447.

[35] A. M. Guess, B. Nyhan, and J. Reifler, "Exposure to untrustworthy websites in the 2016 us election," *Nature human behaviour*, vol. 4, no. 5, pp. 472–480, 2020.

[36] G. Guest, A. Bunce, and L. Johnson, "How many interviews are enough? an experiment with data saturation and variability," *Field methods*, vol. 18, no. 1, pp. 59–82, 2006.

[37] S. S. Guimarães, J. C. S. Reis, F. N. Ribeiro, and F. Benevenuto, "Characterizing Toxicity on Facebook Comments in Brazil," in *Proceedings of the Brazilian Symposium on Multimedia and the Web*, ser. WebMedia '20. New York, NY, USA: Association for Computing Machinery, Nov. 2020, pp. 253–260.

[38] S. S. Guimarães, J. C. S. Reis, M. Vasconcelos, and F. Benevenuto, "Characterizing political bias and comments associated with news on Brazilian Facebook," *Social Network Analysis and Mining*, vol. 11, no. 1, p. 94, Oct. 2021.

[39] T. Hamdi, H. Slimi, I. Bounhas, and Y. Slimani, "A hybrid approach for fake news detection in twitter based on user features and graph embedding," in *International conference on distributed computing and internet technology*. Springer, 2020, pp. 266–280.

[40] Harvard Kennedy School, "Media manipulation casebook," *https://mediamanipulation.org/*, accessed: 2022-02-01.

[41] N. Hassan, F. Arslan, C. Li, and M. Tremayne, "Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 1803–1812.

[42] M. Hobbs, H. Della Bosca, D. Schlosberg, and C. Sun, "Turf wars: Using social media network analysis to examine the suspected astroturfing campaign for the adani carmichael coal mine on twitter," *Journal of public affairs*, vol. 20, no. 2, p. e2057, 2020.

[43] S. Horawalavithana, K. W. Ng, and A. Iamnitchi, "Twitter is the megaphone of cross-platform messaging on the white helmets," in *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 2020, pp. 235–244.

[44] B. Horne and S. Adali, "This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, 2017.

[45] A. Hounsel, J. Holland, B. Kaiser, K. Borgolte, N. Feamster, and J. Mayer, "Identifying disinformation websites using infrastructure features," in *10th USENIX Workshop on Free and Open Communications on the Internet (FOCI 20)*, 2020.

[46] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck, "Automatic detection of generated text is easiest when humans are fooled," *arXiv preprint arXiv:1911.00650*, 2019.

[47] C. Jack, "Lexicon of lies: Terms for problematic information," *Data & Society*, vol. 3, no. 22, pp. 1094–1096, 2017.

[48] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2018, pp. 19–35.

[49] N. Jain, P. Agarwal, and J. Pruthi, "Hashjacker-detection and analysis of hashtag hijacking on twitter," *International journal of computer applications*, vol. 114, no. 19, 2015.

[50] M. Júnior, P. Melo, A. P. C. da Silva, F. Benevenuto, and J. Almeida, "Towards Understanding the Use of Telegram by Political Groups in Brazil," in *Proceedings of the Brazilian Symposium on Multimedia and the Web*, ser. WebMedia '21. New York, NY, USA: Association for Computing Machinery, Nov. 2021, pp. 237–244.

[51] B. Kaiser, J. Wei, E. Lucherini, K. Lee, J. N. Matias, and J. R. Mayer, "Adapting security warnings to counter online disinformation," in *30th USENIX Security Symposium, August 11-13*. USENIX Association, 2021, pp. 1163–1180.

[52] E. Kapantai, A. Christopoulou, C. Berberidis, and V. Peristeras, "A systematic literature review on disinformation: Toward a unified taxonomical framework," *New Media & Society*, vol. 23, no. 5, pp. 1301–1326, 2021.

[53] N. A. Karlova and K. E. Fisher, "A social diffusion model of misinformation and disinformation for understanding human information behaviour." *Information Research*, vol. 18, no. 1, 2013.

[54] F. B. Keller, D. Schoch, S. Stier, and J. Yang, "Political astroturfing on twitter: How to coordinate a disinformation campaign," *Political Communication*, vol. 37, no. 2, pp. 256–280, 2020.

[55] Ł. Korycki and B. Krawczyk, "Adversarial concept drift detection under poisoning attacks for robust data stream mining," *arXiv preprint arXiv:2009.09497*, 2020.

[56] S. Kumar and N. Shah, "False information on web and social media: A survey," *arXiv preprint arXiv:1804.08559*, 2018.

[57] S. Kumar, R. West, and J. Leskovec, "Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes," in *Proceedings of the 25th international conference on World Wide Web*, 2016, pp. 591–602.

[58] M. Latah, "Detection of malicious social bots: A survey and a refined taxonomy," *Expert Systems with Applications*, vol. 151, p. 113383, 2020.

[59] Lockheed Martin, "The cyber kill chain," *https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html*, accessed: 2022-02-01.

[60] L. Luceri, S. Giordano, and E. Ferrara, "Detecting troll behavior via inverse reinforcement learning: A case study of russian trolls in the 2016 us election," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 417–427.

[61] J. Lukito, J. Suk, Y. Zhang, L. Doroshenko, S. J. Kim, M.-H. Su, Y. Xia, D. Freelon, and C. Wells, "The wolves in sheep's clothing: How russia's internet research agency tweets appeared in us news as vox populi," *The International Journal of Press/Politics*, vol. 25, no. 2, pp. 196–216, 2020.

[62] J. Ma, W. Gao, and K.-F. Wong, "Detect rumors in microblog posts using propagation structure via kernel learning," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 708–717.

[63] S. K. Maity, A. Chakraborty, P. Goyal, and A. Mukherjee, "Detection of sockpuppets in social media," in *Companion of the 2017 ACM*

Conference on Computer Supported Cooperative Work and Social Computing, 2017, pp. 243–246.

[64] M. N. Marshall, "Sampling for qualitative research," *Family practice*, vol. 13, no. 6, pp. 522–526, 1996.

[65] P. Mell, K. Scarfone, and S. Romanosky, "Common vulnerability scoring system," *IEEE Security & Privacy*, vol. 4, no. 6, pp. 85–89, 2006.

[66] P. Melo, J. Messias, G. Resende, K. Garimella, J. Almeida, and F. Benevenuto, "WhatsApp Monitor: A Fact-Checking System for WhatsApp," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, pp. 676–677, Jul. 2019.

[67] M. Mendoza, M. Tesconi, and S. Cresci, "Bots in social and interaction networks: detection and impact estimation," *ACM Transactions on Information Systems (TOIS)*, vol. 39, no. 1, pp. 1–32, 2020.

[68] Meta (Facebook), "Coordinated inauthentic behavior archives," *https://about.fb.com/news/tag/coordinated-inauthentic-behavior/*, accessed: 2022-02-01.

[69] P. Mousavi and J. Ouyang, "Detecting hashtag hijacking for hashtag activism," in *Proceedings of the 1st Workshop on NLP for Positive Impact*, 2021, pp. 82–92.

[70] National Institute of Standards and Technology, "Framework for improving critical infrastructure cybersecurity," *https://www.nist.gov/cyberframework*, 2018, accessed: 2022-02-01.

[71] OASIS Cyber Threat Intelligence, "Structured Threat Information Expression (STIX)," *https://oasis-open.github.io/cti-documentation/*, accessed: 2022-02-01.

[72] M. Orabi, D. Mouheb, Z. Al Aghbari, and I. Kamel, "Detection of bots in social media: A systematic review," *Information Processing & Management*, vol. 57, no. 4, p. 102250, 2020.

[73] J. Paavola, T. Helo, H. Jalonen, M. Sartonen, and A.-M. Huhtinen, "Understanding the trolling phenomenon: The automated detection of bots and cyborgs in the social media," *Journal of Information Warfare*, vol. 15, no. 4, pp. 100–111, 2016.

[74] J. Pamment, *The EU's Role in Fighting Disinformation: An EU disinformation framework*, ser. Future Threats, Future Solutions. Carnegie Endowment for International Peace, Sep. 2020, no. 2.

[75] A. Paudice, L. Muñoz-González, A. Gyorgy, and E. C. Lupu, "Detection of adversarial training examples in poisoning attacks through anomaly detection," *arXiv preprint arXiv:1802.03041*, 2018.

[76] J. Peng, S. Detchon, K.-K. R. Choo, and H. Ashman, "Astroturfing detection in social media: a binary n-gram–based approach," *Concurrency and Computation: Practice and Experience*, vol. 29, no. 17, p. e4013, 2017.

[77] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum, "Where the truth lies: Explaining the credibility of emerging claims on the web and social media," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 1003–1012.

[78] N. Rahman, M. Maimuna, A. Begum, M. Ahmed, M. S. Arefin *et al.*, "A survey of data mining techniques in the field of cyborg mining," in *Soft Computing for Security Applications*. Springer, 2022, pp. 781–797.

[79] J. C. S. Reis, P. Melo, K. Garimella, J. M. Almeida, D. Eckles, and F. Benevenuto, "A Dataset of Fact-Checked Images Shared on WhatsApp During the Brazilian and Indian Elections," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, pp. 903–908, May 2020.

[80] L. Reppell and E. Shein, "Disinformation campaigns and hate speech: Exploring the relationship and programming interventions," *Arlington, VA: International Foundation for Electoral Systems*, 2019.

[81] M. H. Ribeiro, S. Zannettou, O. Goga, F. Benevenuto, and R. West, "What do fact checkers fact-check when?" *arXiv preprint arXiv:2109.09322*, 2021.

[82] K. Rieck, T. Holz, C. Willems, P. Düssel, and P. Laskov, "Learning and classification of malware behavior," in *Proceedings of the 5th International Conference Detection of Intrusions and Malware, and Vulnerability Assessment, DIMVA 2008, Paris, France*, ser. Lecture Notes in Computer Science, D. Zamboni, Ed., vol. 5137. Springer, 2008, pp. 108–125.

[83] M. H. Saeed, S. Ali, J. Blackburn, E. De Cristofaro, S. Zannettou, and G. Stringhini, "Trollmagnifier: Detecting state-sponsored troll accounts on reddit," *arXiv preprint arXiv:2112.00443*, 2021.

[84] P. Saha, B. Mathew, K. Garimella, and A. Mukherjee, ""Short is the Road that Leads from Fear to Hate": Fear Speech in Indian WhatsApp Groups," in *Proceedings of the Web Conference 2021*, ser. WWW '21. New York, NY, USA: Association for Computing Machinery, Apr. 2021, pp. 1110–1121.

[85] S. Sardarizadeh and J. Lussenhop, "The 65 days that led to chaos at the capitol," *https://www.bbc.com/news/world-us-canada-55592332*, 2021, accessed: 2022-02-01.

[86] A. Satariano and A. Tsang, "Who's spreading disinformation in U.K. election? you might be surprised," *https://nyti.ms/2YyBhXr*, accessed: 2022-02-01.

[87] M. K. Scheuerman, J. A. Jiang, C. Fiesler, and J. R. Brubaker, "A framework of severity for harmful content online," *Proceedings of the ACM on Human-Computer Interaction*, no. CSCW, pp. 1–33, 2021.

[88] W. Shahid, Y. Li, D. Staples, G. Amin, S. Hakak, and A. Ghorbani, "Are you a cyborg, bot or human?—a survey on detecting fake news spreaders," *IEEE Access*, vol. 10, pp. 27 069–27 083, 2022.

[89] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, "The spread of low-credibility content by social bots," *Nature communications*, vol. 9, no. 1, pp. 1–9, 2018.

[90] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu, "Combating fake news: A survey on identification and mitigation techniques," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 3, apr 2019.

[91] N. Shevchenko, T. A. Chick, P. O'Riordan, T. P. Scanlon, and C. Woody, "Threat modeling: a summary of available methods," *Carnegie Mellon University Software Engineering Institute Digital Library*, 2018.

[92] K. Shu, H. R. Bernard, and H. Liu, "Studying fake news via network analysis: detection and mitigation," in *Emerging research challenges and opportunities in computational social network analysis and mining*. Springer, 2019, pp. 43–65.

[93] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[94] K. Shu, S. Wang, and H. Liu, "Understanding user profiles on social media for fake news detection," in *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE, 2018, pp. 430–435.

[95] M. Silva, L. Santos de Oliveira, A. Andreou, P. O. Vaz de Melo, O. Goga, and F. Benevenuto, "Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook," in *Proceedings of The Web Conference 2020*, ser. WWW '20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 224–234.

[96] K. Starbird, A. Arif, and T. Wilson, "Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, nov 2019.

[97] K. Starbird, A. Arif, T. Wilson, K. Van Koevering, K. Yefimova, and D. Scarnecchia, "Ecosystem or echo-system? exploring content sharing across alternative media domains," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, 2018.

[98] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro, "Some like it hoax: Automated fake news detection in social networks," *arXiv preprint arXiv:1704.07506*, 2017.

[99] E. Taylor, S. Walsh, and S. Bradshaw, "Industry responses to the malicious use of social media," *Nato Stratcom*, 2018.

[100] P. Tehlan, R. Madaan, and K. K. Bhatia, "A spam detection mechamism in social media using soft computing," in *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 2019, pp. 950–955.

[101] The MITRE Corporation, "Common attack pattern enumerations and classifications," *https://capec.mitre.org/*, accessed: 2022-02-01.

[102] ——, "Common weakness enumeration," *https://cwe.mitre.org/*, accessed: 2022-02-01.

[103] ——, "Mitre ATT&CK knowledge base," *https://attack.mitre.org/*, accessed: 2022-02-01.

[104] M. Tomaiuolo, G. Lombardo, M. Mordonini, S. Cagnoni, and A. Poggi, "A survey on troll detection," *Future internet*, vol. 12, no. 2, p. 31, 2020.

[105] Twitter, "Information operations," *https://transparency.twitter.com/en/reports/information-operations.html*, accessed: 2022-02-01.

[106] C. VanDam and P.-N. Tan, "Detecting hashtag hijacking from twitter," in *Proceedings of the 8th ACM Conference on Web Science*, 2016, pp. 370–371.

[107] L. Vargas, P. Emami, and P. Traynor, "On the detection of disinformation campaign activity with network analysis," in *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, 2020, pp. 133–146.

[108] Verizon RISK, "The vocabulary for event recording and incident sharing (VERIS)," *http://veriscommunity.net/incident-desc.html*, accessed: 2022-02-01.

[109] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[110] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng, and B. Y. Zhao, "You are how you click: Clickstream analysis for sybil detection," in *22nd USENIX Security Symposium*, 2013, pp. 241–256.

[111] X. Wang, H. Guo, S. Hu, M.-C. Chang, and S. Lyu, "Gan-generated faces detection: A survey and new perspectives," *arXiv preprint arXiv:2202.07145*, 2022.

[112] Y. Wang, F. Tahmasbi, J. Blackburn, B. Bradlyn, E. D. Cristofaro, D. Magerman, S. Zannettou, and G. Stringhini, "Understanding the use of fauxtography on social media," in *ICWSM*. AAAI Press, 2021, pp. 776–786.

[113] Y. Wang, S. Zannettou, J. Blackburn, B. Bradlyn, E. De Cristofaro, and G. Stringhini, "A multi-platform analysis of political news discussion and sharing on web communities," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 1481–1492.

[114] C. Wardle and H. Derakhshan, "Information disorder: Toward an interdisciplinary framework for research and policy making," *Council of Europe*, vol. 27, 2017.

[115] C. Wardle, H. Derakhshan *et al.*, "Thinking about 'information disorder': formats of misinformation, disinformation, and mal-information," *Ireton, Cherilyn; Posetti, Julie. Journalism,'fake news'& disinformation. Paris: Unesco*, pp. 43–54, 2018.

[116] D. Weber and F. Neumann, "Amplifying influence through coordinated behaviour in social networks," *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 1–42, 2021.

[117] M. Westerlund, "The emergence of deepfake technology: A review," *Technology Innovation Management Review*, vol. 9, pp. 40–53, 11/2019 2019.

[118] T. Wilson and K. Starbird, "Cross-platform disinformation campaigns: lessons learned and next steps," *Harvard Kennedy School Misinformation Review*, vol. 1, no. 1, 2020.

[119] L. Wu and H. Liu, "Tracing fake-news footprints: Characterizing social media messages by how they propagate," in *Proceedings of the eleventh ACM international conference on Web Search and Data Mining*, 2018, pp. 637–645.

[120] Z. Yamak, J. Saunier, and L. Vercouter, "Sockscatch: Automatic detection and grouping of sockpuppets in social media," *Knowledge-Based Systems*, vol. 149, pp. 124–142, 2018.

[121] W. Yaqub, O. Kakhidze, M. L. Brockman, N. D. Memon, and S. Patil, "Effects of credibility indicators on social media news sharing intent," in *CHI'20: CHI Conference on Human Factors in Computing Systems, April 25-30, 2020*. ACM, 2020, pp. 1–14.

[122] S. Zannettou, B. Bradlyn, E. De Cristofaro, G. Stringhini, and J. Blackburn, "Characterizing the use of images by state-sponsored troll accounts on twitter," *arXiv preprint arXiv:1901.05997*, 2019.

[123] S. Zannettou, T. Caulfield, B. Bradlyn, E. D. Cristofaro, G. Stringhini, and J. Blackburn, "Characterizing the use of images in state-sponsored information warfare operations by russian trolls on twitter," in *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM*. AAAI Press, 2020, pp. 774–785.

[124] S. Zannettou, T. Caulfield, E. De Cristofaro, M. Sirivianos, G. Stringhini, and J. Blackburn, "Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web," in *Companion Proceedings of The 2019 World Wide Web Conference*, ser. WWW '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 218–226.

[125] S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini, and J. Blackburn, "Who let the trolls out? towards understanding state-sponsored trolls," in *Proceedings of the 10th acm conference on web science*, 2019, pp. 353–362.

[126] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending against neural fake news," *NeurIPS*, 2020.

[127] A. X. Zhang, A. Ranganathan, S. E. Metz, S. Appling, C. M. Sehat, N. Gilmore, N. B. Adams, E. Vincent, J. Lee, M. Robbins, E. Bice, S. Hawke, D. Karger, and A. X. Mina, "A structured response to misinformation: Defining and annotating credibility indicators in news articles," *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*, pp. 603–612, 4 2018.

[128] Z. Zhao, P. Resnick, and Q. Mei, "Enquiring minds: Early detection of rumors in social media from enquiry posts," in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 1395–1405.

[129] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–40, 2020.

APPENDIX

A. Interview Slide Deck

We provide an abridged version of the questionnaire used to guide our open-ended conversations with the participants.

- **Background: Role/Team/Organization**
  - Describe your role and the different roles within your team with respect to mis-/disinformation.
- **Background: Project(s)**
  - Describe one (or more) projects/events you focused on (Platforms, Coordinatation, Actors, Sophistication)
  - What tools do you use to help you on these projects?
- **Selecting Projects**
  - How do you determine the initial set of projects/events to work on?
  - How do you and your team currently prioritize projects to work on? (Factors, Decision-making process, Tools)
  - What challenges do you face?
- **Assessing Projects**
  - Once chosen for investigation, how do you evaluate a project/event?
  - What are the current processes you use for scoring/labeling a project/event?
  - How do you convey this score/label to your audiences/in your reports?
- **Actor's Motivation and Capabilities**
  - Who are the usual actors behind such events?
  - What are their motivations?
  - What are their capabilities?
    - Amount of control/influence over the platform
    - Level of coordination observed
    - Level of sophistication
- **Wishlist**
  - Setting aside feasibility for a while, what tools/solutions would be most useful for your team to better prioritize projects to focus on and to evaluate a project and assign a score/label?
  - Would you test a tool that assesses the priority of different projects?