The Fragile Balance of Terror
Sagan, Scott D., Narang, Vipin

Published by Cornell University Press

Sagan, Scott D. and Vipin Narang.
The Fragile Balance of Terror: Deterrence in the New Nuclear Age.

For additional information about this book
https://muse.jhu.edu/book/99937

For content related to this chapter
https://muse.jhu.edu/related_content?type=book&id=3295415
Social media platforms such as Twitter, Facebook, and WhatsApp—and their local variants worldwide—have injected new dynamics into the way leaders, governments, and media interact with citizens and vice versa. Leaders and organizations can bypass official bureaucratic channels and communicate, in real time, directly not only with their constituents, but also with anyone in the world with a smartphone or, in the case of open platforms such as Twitter, anyone with access to the internet and a social media account. Crisis tweeting by leaders as a form of communication is a novel feature of the information landscape. During the January 2020 crisis between the United States and Iran, for example, following the assassination of General Qasem Soleimani, President Donald Trump sent 182 tweets, including one on January 4, threatening to target fifty-two Iranian cultural sites.\footnote{Donald Trump, Twitter, January 4, 2020. The full thread reads, “Iran is talking very boldly about targeting certain USA assets as revenge for our ridding the world of their terrorist leader who had just killed an American, & badly wounded many others, not to mention all of the people he had killed over his lifetime, including recently . . . hundreds of Iranian protesters. He was already attacking our Embassy, and preparing for additional hits in other locations. Iran has been nothing but problems for many years. Let this serve as a WARNING that if Iran strikes any Americans, or American assets, we have . . . targeted 52 Iranian sites (representing the 52 American hostages taken by Iran many years ago), some at a very high level & important to Iran & the Iranian culture, and those targets, and Iran itself, WILL BE HIT VERY FAST AND VERY HARD. The USA wants no more threats!”} In addition to serving as a unique communications platform for leaders, ordinary citizens can interact with, disseminate, like, ridicule, or otherwise reach leaders and organizations that they previously could never dream of directly interacting with—and not just celebrities, but their elected (or unelected) leaders.

How have these platforms affected international politics, particularly crises involving one or more nuclear powers? Do they trigger crises and serve
as rocket fuel for their escalation? Or are they just noise that thickens the fog of war without fundamentally altering crisis dynamics? And how do different platforms affect crisis dynamics as the sheer volume of information—some of which is of dubious quality and produced by an unsubstantiated rumor mill—pollinates and pollutes the media and popular discourse?

These are questions on which we have little scholarship or thinking in international security. This chapter attempts to fill this gap by disaggregating social media platforms along three relevant dimensions in crises—whether they are open or closed platforms (platform properties), whether the event is discrete or ongoing (crisis properties), and whether the intended audience is international or domestic (audience properties). We provide evidence from South Asia, notably the 2019 Pulwama/Balakot crisis, the 2018 Hawaii missile alert, and the 2017 fake evacuation order on the Korean Peninsula, to illustrate how and whether social media platforms affect crisis dynamics.

We hypothesize and provide preliminary evidence that social media does not have a uniform effect on crisis dynamics but can have varying impacts depending on the specific type of platform and the nature of the crisis. In particular, open social media platforms such as Twitter have the potential to thicken the fog of war in discrete crises. But open platforms are also more likely to assemble and disseminate accurate information more broadly in ongoing crises. Closed platforms such as WhatsApp and Facebook may reinforce incorrect information even over protracted crises, as users gravitate toward social clusters that reinforce preexisting beliefs and opinions rather than seeking accurate information. Open platforms may also be more useful for external signaling to adversaries and a broader international community; their wide reach allows for deterrence or coercive signals to be more quickly and directly transmitted to target audiences. Closed platforms, by contrast, can be exploited to generate nationalism or domestic audience costs by specifically and privately targeting and potentially monopolizing the information constituents receive. All these platforms are also vulnerable to manipulation either by governments to generate domestic support or by malicious actors that can employ them to spread potentially dangerous disinformation.

Do social media platforms affect the outbreak or course of nuclear crises? For example, would the existence of Twitter or WhatsApp during the 1999 Kargil war or 2008 Mumbai attack between India and Pakistan have led to a different outcome, namely escalation or even nuclear use? The media environment during Kargil, or during the 2008 terrorist attack in Mumbai, was riddled with substantial misinformation. Ultimately, Kargil did not escalate

---

because India’s prime minister Atal Bihari Vajpayee did not want to escalate and ordered a highly constrained Indian response to the Pakistani infiltration. It is hard to envision that if live streaming television of these events in both Hindi and English at the time did not generate pressure for escalation, Twitter would have. In another counterfactual, if social media did not exist during the 2018 Hawaii missile alert, traditional media and communication platforms, such as television, radio, and text messaging, might have achieved the same outcome and de-escalated a potential crisis. Although social media platforms may not independently cause crises or escalation, they can contribute to the broader information and misinformation environment, notably by cross-pollinating (or cross-polluting) more mainstream media platforms such as television or print news. They can also be manipulated by governments and malicious actors to affect crisis dynamics, serving as accelerants or amplifiers, with both risks and benefits. They provide a mechanism to de-escalate crises quickly and serve as a propellant to spin them up, with the risk of nuclear powers tweeting their way to war.

Social media platforms as a whole inject a level of complexity and information velocity into nuclear-tinged crises that nuclear weapons powers have not previously had to manage. Our overall conclusion is a simple but underappreciated point thus far: different social media platforms can work at cross-purposes against each other in crises, making a uniform effect difficult to measure or even theorize. In aggregate, the new social media ecosystem may simply generate noise and ambiguity, which can sometimes make navigating a crisis difficult but may also provide states face-saving ways to de-escalate. Each platform may have differential effects depending on the type and duration of the crisis, as well as which audience is being targeted. This chapter undertakes a brush-clearing exercise to systematically think about how different social media platforms may impact different types of nuclear crises. We disaggregate social media platforms—since social media is not a monolithic concept or tool—and crisis types to theorize about the different ways in which social media may affect crises between nuclear actors. These dynamics may apply to all types of militarized crises between states, but we restrict our focus to crises with a nuclear component, given the risks involved with rapid or unintended escalation in such crises.


4. See, for example, Harold A. Trinkunas, Herbert S. Lin, and Benjamin Loehrke, Three Tweets to Midnight: Effects of the Global Information Ecosystem on the Risk of Nuclear Conflict (Stanford, CA: Hoover Institution Press, 2020), which offers a variety of excellent perspectives, many of which treat social media platforms as monolithic.

Social Media and Escalation Risks: Existing Pathways

Escalation is ultimately a political activity. Although it is typically thought of as a linear process, “the sequential expansion of the scope or intensity of conflict,” the increasing speed of conflict and emergence of new technologies has complicated the notion of an “escalation ladder” and replaced it with something more akin to an “escalation web.”

Escalation is typically categorized as being either deliberate, such as a preemptive nuclear attack, or in the face of defeat; or inadvertent, to include accidents, unintended use of force from apex political authorities, “mechanical failure, unauthorized (nuclear) use, or insanity.” The election of US president Trump, accompanied by his novel approach to tweeting, brought attention to the intersection of social media and politics, particularly nuclear weapons postures—whose button is really bigger?—and nuclear threats such as “fire and fury.” But Trump is hardly the only world leader to use social media in an inflammatory way, and in 2019 Twitter issued guidelines for “World Leaders: Principles and approaches” warning them against tweets that could promote terrorism, threaten violence, encourage self-harm, or risk an enforcement response by Twitter.

For violations of these guidelines and due to the risk of further incitement of violence in the United States, Twitter suspended President Trump’s account—terminating his ability to tweet—shortly after the events of January 6, 2021, with two weeks left in office, an unprecedented move.

At its core, social media is a means of virtually connecting people in real time with limited, if any restrictions, on content. The most common social media platforms include Facebook, YouTube, WhatsApp, Instagram, and Twitter. Platforms are differentiated by a variety of factors: accessibility of information (open or closed), censorship and government oversight, content (visuals, videos, or text), sponsorship, and human versus bot-generated content, among others. Social media is increasingly present in international politics and conflicts, such as the 2014 Ukraine crisis with activists.

---


9. Statista, “Most Popular Social Networks Worldwide as of April 2020, Ranked by Number of Users,” https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. Interestingly, although Twitter is one of the most commonly studied platforms with regards to politics, it is only the twelfth most popular social media platform, behind QZone, TikTok, and Reddit.
experts, politicians, and even armed fighters taking to platforms such as Twitter, Facebook, and YouTube to document and circulate their experiences in real time with multimedia tools.

Scholarship into the impact of social media on politics and nuclear weapons, specifically, is relatively nascent. A handful of studies across disciplines highlight at least four related ways social media may have an escalatory effect in crises, and their combination may be both figurative and literal confusion: disinformation, ambiguity, impact on decision making, and public pressure. First, social media is a primary vehicle for disseminating outright disinformation. Lanoszka defines disinformation as “a systematic government effort aimed at using disinformation to mislead a particular audience—whether a government or key members of society—in order to influence the policy process.”[^10] On social media, the disinformation need not originate from the government; in some cases the government may propagate it, but in others, the government may be embattled to correct or fight disinformation campaigns that originate from malicious nonstate actors or influencers. Disinformation can range from rumors to deep fakes. A 2018 study by MIT researchers based on 126,000 stories tweeted by 3 million people between 2006 and 2017 found that “Falsehood diffused significantly farther, faster, deeper, and more broadly than truth in all categories of information, and the effects were more pronounced for false political news than for false news about terrorism, natural disasters, science, urban legends, or financial information.”[^11] Interestingly, the study also found that it took the truth six times as long as the falsehood to reach 1,500 people, perhaps confirming the old adage that the truth is boring, while disinformation is often spectacular and shocking.[^12]

It remains unclear if this wider and faster dissemination of disinformation has a significant impact on politics. Lanoszka demonstrates that “disinformation is ineffective in terms of changing the policies of a target as regards to its foreign policy alignments and armaments—that is, the balance of power.”[^13] Rather than create new tensions and pressure escalation, rumor typically reinforces preexisting views.[^14] Of particular concern is the spread of disinformation during a crisis, when rumors have been shown to flourish and propagate virally.[^15]

A second concern is that social media platforms may contribute to unintended escalation due to enhanced ambiguity, or the thickening of the fog of war where leaders or governments are less certain as to the facts on the ground and are forced to assume the worst, erring on the side of escalation. Posen explored inadvertent escalation scenarios during the Cold War that involved targeting nuclear systems due to the inability to discriminate facts and systems in real time.\textsuperscript{16} Social media might provide information in real time; however, this is not necessarily accurate information, nor can it help differentiate offensive from defensive acts.\textsuperscript{17} By flooding the information space, sowing uncertainty, and generating mixed messages in the midst of an already chaotic crisis environment, social media certainly can have escalatory consequences as states assume the worst about the situation or their adversary in the face of ambiguity.\textsuperscript{18}

Current scholarship is mixed on whether ambiguous messaging has a stabilizing or destabilizing effect during crises. On the one hand, it may force states to assume the worst-case scenario and overreact; on the other hand, it may allow for face-saving pathways to de-escalation. The conventional wisdom tends to support the former, arguing that in the absence of reliable or consistent information, actors may misperceive an adversary’s capabilities or willingness to respond.\textsuperscript{19} Misperception occurs when there is an unintentional discrepancy between a state’s actions and the adversary’s understanding of those actions.\textsuperscript{20} By potentially injecting more noise into the system, social media may increase the propensity that such a discrepancy will not only exist but be larger than in the past. Although deliberate ambiguity is central to nuclear deterrence theory, such as Thomas Schelling’s “threat that leaves something to chance,” social media amplifies the risk that states perceive a threat when none was made, or vice versa.

Inherent within this concern about unintentional escalation is the impact of social media on decision making, a third theme in the emerging literature on social media. A 2020 edited volume identifies the impact on decision-making calculus as the “most worrisome possibility” due to the spread of disinformation.\textsuperscript{21} Malicious actors—whether a state actor, state-supported nonstate actor,

\textsuperscript{17} Posen, “Inadvertent Nuclear War?,” 35.
\textsuperscript{20} Daniel S. Geller, “Nuclear Weapons, Deterrence, and Crisis Escalation,” \textit{Journal of Conflict Resolution} 34, no. 2 (June 1990), 293.
\textsuperscript{21} Trinkunas, Lin, and Loehrke, \textit{Three Tweets to Midnight}, 4.
or independent actor—can manipulate social media to distort messaging throughout an ongoing crisis. At the onset of a crisis, in particular, social media may increase pressure on decision makers to “do something,” whereas silence on social media could prompt worst-case thinking and preemptive attacks.\textsuperscript{22} But if the conflict escalates to a limited war, Lawrence Freedman identifies increasing complexity confronting decision makers as they are forced to account for a rich mix of factors: military logic, commitments, international law, domestic politics, casualty levels, economics, and dangers of a wider conflict.\textsuperscript{23} But what if images—or fakes—shared via Twitter or Instagram suggest a higher casualty level than official figures? How might this impact a leader’s perception of the crisis, absent definitive intelligence? These questions are exacerbated when the stakes (and emotions) are higher, and social media may complicate opportunities for dialogue and interaction among political leaders.\textsuperscript{24}

It is possible that social media does not necessarily change crisis outcomes or significantly change leaders’ preferences about escalation. A historical counterfactual may be illustrative, such as if John F. Kennedy and Nikita Khrushchev had the ability to tweet during the Cuban Missile Crisis. According to Freedman, “in Cuba the drive to escalation was inhibited by keen awareness on both sides that the political stakes were simply not worth taking matters to a decisive showdown.”\textsuperscript{25} Unless social media affects underlying political stakes, it is not obvious that it will change crisis behavior. Indeed, the Cuban Missile Crisis counterfactual has to also account for the possibility that Kennedy could have just as easily tweeted de-escalatory messages to more rapidly end the crisis: “I want the missiles out, but I do not want a war Nikita,” with Khrushchev replying back, “Me neither Jack, let’s talk.” Therefore, the pathways to rapid escalation imbued within social media in crises allow for equally rapid de-escalation if both sides have a common preference for it.

A fourth and final theme is the concern that social media might amplify nationalism and public pressure for escalation. Social media could influence escalation dynamics by negatively shaping public opinion, largely through rumors and increasing calls for deliberate escalation to preempt an anticipated attack or stop a costly conflict. Social rumor can be defined as “improvised and expressive forms of public opinion that especially arise under

\textsuperscript{22} Herbert Lin, “Escalation Dynamics and Conflict Termination in Cyberspace,” \textit{Strategic Studies Quarterly} 6, no. 3 (Fall 2012): 58.
\textsuperscript{24} Paul Slovic and Herbert S. Lin, “The Caveman and the Bomb in the Digital Age,” in \textit{Three Tweets to Midnight}, 40.
\textsuperscript{25} Freedman, “Escalators,” 27.
uncertain socio-political situations” and has been proven to exacerbate conflict.\textsuperscript{26} In addition, research demonstrated that visual images and videos are five times as likely to be shared.\textsuperscript{27} In a crisis that the public can see, Instagram may prove more useful than Twitter—and together they can be, quite literally, deadly. And rapid (private or public) dissemination of accurate or fake images/videos as opposed to just text descriptions of the same in a crisis—especially of death or suffering—can mobilize emotions and revenge motives in ways that generate escalatory pressures.\textsuperscript{28} Research into non-nuclear crises demonstrates that whether publics believe a rumor on social media “is determined by pre-existing cognitive schemes and attitudes rather than simply by credulity or gullibility.”\textsuperscript{29} For the most part, publics use social media during a crisis to reduce uncertainty by verifying information and disproving false rumors, foster a shared “keynote” narrative of events, and most—but not all—emergency information on Twitter was found to be accurate, and audiences can largely identify and dismiss untrue rumors.\textsuperscript{30}

A confounder in any analysis of social media dynamics is cross-pollination with mainstream media since many journalists and television outlets now derive information from, as well as break news on, platforms such as Twitter. This cross-pollination dynamic can serve as rocket fuel for misinformation and rumormongering, spreading misinformation to much wider audiences and making it even more difficult to correct. Mainstream media and its relationship to social media outlets means all are competing for breaking news, and seemingly credible, but sensational tweets often cross-pollinate the mainstream media, amplifying their reach and impact. In many ways the current impact of social media is comparable to the “CNN effect” and the advent of the 24/7 news cycle in the late 1980s and early 1990s. Arguably, the CNN effect increased public pressure on governments to shift their foreign policies, particularly military interventions based on what people were seeing—visually in ways they had not previously—at home. This might entail sending humanitarian aid to Somalia or drawing down operations in Afghanistan with visuals of body counts. Where social media differs from the CNN effect, of course, is that in the current climate, governments are just as subject to the “Twitter effect” as the public—they are


\textsuperscript{27} See, for example, Williams and Drew, “Escalation by Tweet.”


\textsuperscript{29} Kwon et al., “Social Media,” 5.

\textsuperscript{30} Tanja Schreiner, “Information, Opinion, or Rumor? The Role of Twitter During the Post-Electoral Crisis in Cote D’Ivoire,” \textit{Social Media and Society} (January–March 2018), 1–16; Kwon et al., “Social Media,” 4.
reading the same messages and might be victim to the same disinformation campaigns.

This review of contemporary scholarship also reveals gaps and points of tension in understanding the impact of social media on crises. The majority of scholarship treats social media as monolithic or focuses on a single platform, with no differentiation between platforms such as Twitter or TikTok. Similarly, the social media literature fails to disaggregate the impact of social media on short, sharp crises versus longer ones. We hypothesize that not only does the type of social media platform matter for escalation dynamics, but the type of crisis matters as well. A significant reason for the confusion within the literature on the impact of social media on crises is that it treats all social media platforms and all crises as uniform. We aim to correct these mistaken assumptions by disaggregating both social media and crises, hypothesizing that different types of social media platforms can have varied effects in different types of crises.

**Toward a More Fine-Grained Treatment of Thermonuclear Twitter**

This chapter offers an original framework to systematically analyze the impact of social media on escalation based on the properties of three dimensions: the type of platform, the type of crisis, and the target audience. Our goal is to offer a useful framework for thinking about the impact of social media on crises in a more granular fashion, which can be applied across platforms, regions, and audiences.

First, the type of platform matters. Namely, is the platform open or closed? We treat this as a binary variable, but it is admittedly potentially continuous, as Twitter has completely open reach—even despite echo chambers—whereas Facebook has potentially several nodes of reach from the user, and text message apps such as WhatsApp can have large private groups or single users. Open platforms, such as Twitter, are often an opportunity to engage with strangers, whereas closed platforms, such as WhatsApp, are how we communicate with those we know and trust. As demonstrated in the study by Hunt Allcott, Matthew Gentzkow, and Chuan Yu, disinformation is more likely to

---

31. An exception to this is Hunt Allcott, Matthew Gentzkow, and Chuan Yu, “Trends in the Diffusion of Misinformation on Social Media,” *Research and Politics* (2019): 108. They analyzed 569 cases of fake news websites and 9,540 fake stories on Facebook and Twitter between January 2015 and July 2018 and found that misinformation had a limited impact over Facebook compared to Twitter—one explanation for this was changes to the Facebook platform after the 2016 US presidential election.

be spread via Twitter than Facebook. But it is theoretically possible that open platforms like Twitter more easily “revert to the mean” over time as disinformation can be publicly corrected in ways that closed platforms such as Facebook and WhatsApp may not enable. In general, we hypothesize that Twitter users seek, and eventually receive, more accurate information in crises as misinformation is corrected—even if dissemination is imperfect—whereas closed platforms are more likely to have misinformation persist and be deployed to reinforce preexisting beliefs, such as hypernationalism.

Second, the nature of the crisis also matters. This is supported by literature on escalation, whereby the political context and stakes affect the outcome of the conflict. Studies on the impact of social media from different disciplines also reinforce this point. A study on Twitter usage during a political crisis in Côte d’Ivoire in 2010–2011, for example, found that social media usage and its impact shifted with the nature of the conflict. At the outset of the conflict, it was primarily used for political signaling for domestic audiences, but as the conflict escalated it shifted to being a vehicle for providing information about humanitarian aid to the public along with sharing information about the nature of the conflict with international audiences. We hypothesize that crisis duration is the relevant property involving the differential impact of social media. We differentiate between short crises ranging from hours to several days, and long crises going beyond that. In short crises focused on a discrete event, such as a singular terrorist attack or explosion (such as the August 2020 ammonium nitrate explosion in Beirut), we hypothesize that misinformation will spread rapidly as “hot takes” make the rounds on both open and closed platforms. But in long crises, involving iterated and longer interaction between adversaries, open platforms are more likely to revert to the mean and widely disseminate more accurate (not necessarily entirely accurate, but a relatively more accurate version) of events than closed platforms, where conspiracy theories and misinformation are likely to persist. We differentiate crises as either discrete or ongoing, but obviously there is a spectrum of duration and intensity, though we hypothesize that duration is the more relevant variable as information reverts to the mean.

The third dimension we focus on is domestic versus international audiences. Whether or not a social media message has an impact begs the question: impact on whom? Scholarship to date focuses on two audiences: international actors, particularly political leaders and decision makers, and the domestic public. Russian president Vladimir Putin is likely to be impacted by a Trump tweet differently than a Trump voter in Florida. That said, there will likely be spillover between domestic and international au-

34. See Bell and MacDonald, “Escalation.”
Table 3.1 Short Crisis

<table>
<thead>
<tr>
<th></th>
<th>Domestic</th>
<th>International</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open (Twitter)</td>
<td>Hot takes/misinformation</td>
<td>Alerting international actors (risk of spillover)</td>
</tr>
<tr>
<td>Closed (WhatsApp)</td>
<td>Fearmongering</td>
<td>Diasporic nationalism?</td>
</tr>
</tbody>
</table>

Table 3.2 Long Crisis

<table>
<thead>
<tr>
<th></th>
<th>Domestic</th>
<th>International</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open (Twitter)</td>
<td>Information improves</td>
<td>Deterrence/coercion signaling (risk of spillover)</td>
</tr>
<tr>
<td>Closed (WhatsApp)</td>
<td>Preexisting beliefs/conspiracy theories</td>
<td>Diasporic patriotism?</td>
</tr>
</tbody>
</table>

diences on open platforms, as the messages targeted at one group will also be read by the other one. In general, social media in crises involve communication, signals, and spillover between leaders to leaders, leaders to the public, the public to leaders, and the public to each other. And, in some cases, messages intended for domestic consumption can have unintentional consequences if misinterpreted by the leaders of the adversary.

Based on these variables, we identify two main sets of hypotheses about how social media impacts nuclear crises. We orient these dimensions in tables 3.1 and 3.2 to generate hypothesized effects based on these three properties of nuclear crises in an era of social media.

- **H1**: Social media does not have a uniform effect on crisis dynamics but can have varying impacts depending on the specific type of platform and the nature of the crisis.
  - H1a: Open social media platforms are more likely to thicken the fog of war than closed platforms during short crises.
  - H1b: Closed social media platforms will reinforce existing views over long crises to include conspiracy theories.
- **H2**: Social media does not have a uniform effect on international and domestic audiences.
  - H2a: Open platforms are more useful for signaling to the international community than closed platforms.
  - H2b: Closed platforms are more likely to be used to exacerbate domestic public opinion and nationalism during a crisis.

Depending on the properties of the platform, the crisis, and the audience, we hypothesize differential effects and dynamics in crises. The aim of this theoretical framework is to offer more fine-grained predictions and mechanisms for how different social media tools affect nuclear crisis dynamics, rather than simply assuming that there is an overall effect and that all platforms have a uniform effect. The dissatisfaction is that there may not be an overall effect, but we believe it is more accurate to treat the impact of social media on crisis dynamics at a more granular, platform level than search for aggregate effects, which may be difficult to identify and isolate.

To explore these hypotheses and illustrate these mechanisms, we examine three empirical cases: the 2019 India-Pakistan crisis, the 2018 Hawaii missile alert, and the 2017 United States Forces Korea (USFK) evacuation order from the Korean Peninsula. The case studies were selected because they all have a nuclear component involving two nuclear actors, at least in the background, and otherwise offer useful variation. The Pulwama/Balakot crisis was a discrete crisis on the day of the attack on Indian security services on February 14, 2019, followed by a several-day crisis when India retaliated on February 26, 2019. The 2018 Hawaii missile alert originated with a push notification through cellular carriers but not via a public platform, thus it almost uniquely had the properties of a universal private network. All cell phone users received an official emergency alert, which was inadvertently sent during a test warning of an incoming North Korean ballistic missile, closing with “This is not a drill.” This generated understandable panic. Although there was no reason to otherwise believe North Korea would launch a first strike against Hawaii, the message was disseminated through an official state channel six weeks after North Korea tested its long-range missile, the Hwasong-15 ICBM, and before Kim Jong Un initiated the so-called charm offensive. And just four months earlier in 2017, at the height of tensions between the United States and North Korea, a fake message ordering the evacuation of nonessential US military personnel from South Korea—the leading indicator of a potential US-led attack against North Korea—circulated on closed platforms such as WhatsApp and Facebook, generating fears of an impending war at a time when it was plausible that the United States may initiate a surprise attack on Kim Jong Un’s North Korea.37 Within the cases, we look for indicators of social media impact, to include (1) responses from international actors over social media; (2) spread of misinformation, disinformation, and “deep fakes”; (3) shifts in public opinion and attitudes; and (4) military responses and escalatory (or de-escalatory) actions.

On February 14, 2019, a local twenty-two-year-old Kashmir man, Adil Ahmad Dar, drove a minivan laden with RDX into a bus carrying paramilitary forces from the Central Reserve Police Force (CRPF) near the Kashmir village of Pulwama, killing over forty personnel in the worst terrorist attack in Indian Kashmir in decades. Within hours, the attack was claimed by the Pakistani militant group, the Jaish-e-Mohammed (JeM), which radicalized Dar and purportedly assisted him with logistics and planning. Indian social media went into hyperdrive. After initial ambiguity about the nature of the attack, India’s major media figures on Twitter arrived at a relatively accurate accounting of what happened: the JeM took responsibility for an attack in Kashmir that killed about forty (the final count ended up being forty-two) CRPF personnel. Initial accounts underreported the number killed, rather than exaggerating them, so the figure circulating on Twitter by the end of the day had been revised up from “over a dozen” to the nearly accurate forty CRPF soldiers. This seems to have been consistent in both English and Hindi language accounts and is consistent with what we expect with open platforms in a relatively discrete crisis, where confusion eventually leads to convergence through crowdsourcing and viral propagation of more updated information.

The dynamics on WhatsApp appear to have been different and illustrate the perils of closed social media platforms during discrete crises. In closed WhatsApp groups a variety of misinformation and disinformation campaigns were circulating virulently and virally. For example, a fake picture of India’s main opposition candidate, Congress’s Rahul Gandhi, who was about to run a national election against the wildly popular BJP incumbent prime minister, Narendra Modi, was photoshopped as meeting Dar, the suicide terrorist. It is unclear how widely this was propagated or believed, as this is difficult to measure. But it was making the rounds on highly popular and widely subscribed BJP linked WhatsApp groups. Older videos, which will only get more realistic with so-called deep fake technology, were circulating that appeared to show Congress celebrating the attack and shouting “Pakistan zindabad” (long live Pakistan) as part of a concerted political effort to

---


paint the Congress Party as Pakistan sympathizers. What was not fake was a message that seems to have circulated millions of times in India on various WhatsApp groups calling for India to destroy Pakistan, fueling hypernationalism and warmongering. Citing each previous war that Pakistan “started” and India “finished,” the message closed with “2019: They Started, and we will BLOODY HELL FINISH IT.” Others demanded 400 Pakistanis killed for 40 Indians, still others demanded that Modi do a “Gujarat” on Pakistan (i.e., massacre Muslims). These were shared virally in multiple groups in both English and Hindi, and all aimed at seemingly one thing: whipping up nationalism and revenge emotions against Pakistan and, dangerously, against Muslims within India. Some were preceded with the hashtag #IndiaNeedsNuclearStrike. Equally worrying were the messages going viral that purported to show the Muslims of India celebrating the attack—creating the real possibility of internal violence and reprisals against Indian Muslims, which did happen sporadically but was thankfully not widespread.

Although Twitter had its fair share of rabble-rousing, the public and open nature of the platform tended to moderate it compared to WhatsApp. The latter, in addition to being a closed network, is also perceived to be a more private space, which has led people to share and say what they would otherwise not say publicly, revealing a widespread virulent nationalism and othering of Pakistan and Indian Muslims. Facebook was also a platform for particularly nationalist comments, and an NDTV editor was suspended for posting racist comments on her Facebook page. Twitter at least allows for a possible pathway for users to be exposed to alternative information and perspectives since the nature of retweets and the platform tends to revert to the mean over time. To be sure, this effect will be moderated by the echo chamber effect—it is those accounts in extreme networks that have the lowest probability of receiving moderating information. But on WhatsApp, such a pathway may not even exist; a user’s potential exposure is limited by the groups that they are in. As such, it is easy to select into groups that reinforce preexisting beliefs, where conspiracy theories and virulent narratives take hold and are difficult to dislodge.

Twelve days later, India struck back at a JeM Madrassa outside a Pakistani town, Balakot. For the first time in history, a nuclear weapons power used airpower directly on the undisputed sovereign territory of another nuclear

42. Purohit, “After Pulwama.”
ThERMONuClEar TwiTTEr?

Interestingly, the first inkling of the attack came from the Twitter account of the official Pakistan Army spokesman, Major General Asif Ghafoor, thereby alerting the world via social media that India had retaliated.\(^44\) For India it was the first time that it had used airpower against mainland Pakistan since the 1971 war. It is probably unsustainable to argue that social media put pressure on Modi to hit back in this fashion since, with or without Twitter or WhatsApp, Modi seems to have been itching to retaliate. This is consistent with our general hypothesis that social media pressure is endogenous to what states may otherwise prefer to do. What social media likely did was prepare the domestic political battlefield to overwhelmingly support a relatively aggressive retaliation. The BJP IT cell went into hyper-drive after the attack, propagating both bot-driven tweets and WhatsApp messages hailing Modi’s retaliation—with his fifty-six-inch chest—as historic after Congress’s failure to do anything, especially after the 2008 Mumbai attack that killed 173 civilians. Immediately, the news and bot-driven tweets (distinguishable by their verbatim language) were deployed on Twitter to claim that Modi had killed 500 or 600 terrorists at Balakot—very close to the 400 deaths demanded by nationalist WhatsApp groups after Pulwama. The same purportedly went out over the WhatsApp groups.

The problem for both the Indian and Pakistani governments was that, at least on Twitter, the global analytical community was watching, analyzing, and weighing in. The Pakistani claim that India’s Air Force turned around after being intercepted was quickly debunked when the official army spokesman later tweeted pictures of Indian missiles hitting trees deep into Pakistani territory, roughly 60 kilometers from the Line of Control (LoC).\(^45\) This was both for domestic political purposes in Pakistan—the Indian Air Force (IAF) did not hit anything except some trees and maybe a goat—but was also the first report to the international community that India had retaliated. On February 26, Major General Ghafoor announced in an official Pakistan Army press conference that was live-streamed, live-tweeted, and later put on YouTube that Pakistan was convening the National Command Authority (NCA), ominously adding—cutting from Urdu to English for maximum international signaling effect—“I hope you understand what is National Command Authority, what does it constitute.”\(^46\) The NCA is the body in charge of Pakistan’s nuclear weapons. This was clearly a signal to

India, but more important, the United States, whose intervention in South Asian crises Pakistan has routinely attempted to catalyze in order to prevent Indian escalation.\textsuperscript{47}

However, the Indian claim that it brought down the Balakot structure was also quickly debunked because it was still standing, and every commercial satellite could still see it, a new feature Zegart explores in her chapter. Within a couple of days, India’s Air Force claimed it used a penetrator version of the Spice standoff missiles that were not designed to bring down the structure. Commercial satellite imagery detected several ground scars on the side of the hill where the structure stood, leading to speculation that most, if not all, of the Spices missed the target. It is possible one or more penetrated the building, but it became obvious that the BJP’s initial claim of several hundred terrorists killed (which was later sourced to how many cell phones the National Technical Research Organization [NTRO] detected several days before the strike, not on the day of the strike) was likely inaccurate and exaggerated.\textsuperscript{48} On Twitter, open-source analysts eventually concluded that the IAF either completely missed, or largely missed, the target at Balakot. But this was mostly in English. Major Hindi language accounts did not bother correcting the original estimate.\textsuperscript{49} On WhatsApp and closed platforms, however, this outside analysis did not find its way into the narrative, and the “man on the street” across India continues to largely believe that Modi killed hundreds of terrorists at Balakot.\textsuperscript{50}

On February 27, the Pakistan Air Force (PAF) retaliated for India’s Balakot strike by crossing the LoC and, it claims, “intentionally missing” an Indian Brigade Headquarters, just to demonstrate that it could hit it if so desired. In the ensuing dogfight, India shot down one of its own Mi-17 helicopters in a tragic friendly fire incident and lost a MiG-21 Bison that had crossed the LoC in hot pursuit of the PAF. The pilot, Wing Commander Abhinandan Varthaman, was captured alive by Pakistan, whose official media spokesman tweeted that there were multiple Indian pilots in custody but then clarified that there was only one.\textsuperscript{51} He was wounded but then expeditiously returned to India on March 1 as a de-escalatory measure. Var-

\textsuperscript{47} See Narang, Nuclear Strategy in the Modern Era.

\textsuperscript{48} “Rajnath Singh Says NTRO Surveillance of JeM Camp in Balakot before IAF Air Strikes Confirmed 300 Active Mobile Phones,” Press Trust of India, March 5, 2019.

\textsuperscript{49} See, for example, ABP News, Twitter, February 27, 2019, https://twitter.com/ABPNews/status/1100765012572815360?s=20.

\textsuperscript{50} See, for example, Pooja Chaudhuri, “Fake WhatsApp Chat Shared on Social Media to Claim 292 Terrorists Killed in Balakot Airstrike,” alt news, March 5, 2019, https://www.altnews.in/fake-whatsapp-chat-shared-on-social-media-to-claim-292-terrorists-killed-in-balakot-airstrike/.

thaman then claimed that he shot down a PAF F-16 before he went down. Thus began one of the great social media mysteries of the Balakot crisis: was there actually a second pilot, or #doosraBanda, as Pakistan’s official military spokesman initially claimed, and if so was it a Pakistani Air Force pilot who ejected from an F-16 that was shot down? Reputable Indian journalists continued to tweet about the #doosraBanda months later, and the IAF offered official briefings showing circumstantial evidence the F-16 was shot down. Pakistan continues to deny that an F-16 pilot was killed and is bolstered by a US Department of Defense (DoD) leak that all US-origin (meaning all) PAF F-16s were later accounted for and operational. WhatsApp groups circulated fake videos of a PAF parachute, and the belief that India shot down an F-16 persists on closed social media platforms. Many Indian journalists and military on open platforms such as Twitter continue to believe so. This debate spilled over into mainstream media outlets, illustrating the cross-pollination effect of social media to mainstream media, which would reproduce tweets and potentially doctored images on-screen during live broadcasts. Much of the WhatsApp activity was designed to whip up patriotism and pride in India’s Armed Forces after spinning up the hypernationalism before the Balakot strike. Anecdotal evidence suggests that almost all Indians adamantly believe that Varthaman killed a PAF F-16, because it aligns with preexisting beliefs in the superiority of India and its Air Force, and that the DoD audit was motivated by Lockheed’s refusal to admit that an F-16 could be killed by a vintage MiG-21 Bison.

A curious episode then emerged after the crisis seemingly de-escalated. Reports surfaced that India may have threatened to escalate the crisis with surface-to-surface missile strikes if Abhinandan had not been expeditiously returned by Pakistan. Pakistan allegedly threatened privately to retaliate “three times over” if India breached that red line. During the campaign trail, Prime Minister Modi seemingly confirmed this threat when he stated that he was prepared to order a missile strike that would have been a “qatal ki raat,” or night of murder. But none of this occurred through social media, suggesting that significant deterrence threats between governments can still be conveyed privately, the old-fashioned way, and not make its way to social media. When this threat was made public, it was for domestic political purposes for

52. See, for example, Vishnu Som, Twitter, October 8, 2019, https://twitter.com/VishnuNDTV/status/1181495301456023537?s=20.
Modi’s reelection campaign to illustrate his resolve and aggressiveness against fighting terrorism.

What can we learn from the social media dynamics during the Pulwama/Balakot crisis in February 2019? For one, different platforms had different dynamics. Open platforms such as Twitter served largely, over time, to generate relatively accurate information for those who wished to find it. There is no doubt a lot of trolling and nationalism on Twitter, but basic facts tended to revert to their true values over time and correct initial misinformation. Here the audiences were domestic and international, for both deterrence signaling, and making the international community aware of events—Pakistan’s official army spokesperson was the first to alert the world to India’s strike at Balakot and mention, in English, Pakistan’s convening of the NCA, perhaps as a signal to catalyze international efforts to intervene in the crisis. Closed groups, such as WhatsApp, were a completely different story and aimed primarily at domestic audiences. Coordinated groups and messages attempted to whip up hypernationalism following the Pulwama attack, and then circulated government propaganda after the Balakot strike and retaliation, with little if any effort by anyone to verify information. This may have largely served to reinforce preexisting beliefs. In a country where a majority of respondents, according to Sagan and Valentino, would have no problem using nuclear weapons first even if there is no military advantage, those preexisting beliefs provide a domestic political base to support quite aggressive policies. This mechanism is potentially escalatory.

But there was a de-escalatory mechanism provided by social media platforms as well. Both sides could convey narratives to their publics that allowed for them to walk away from the crisis claiming they won. India convinced its public that it killed over 300 terrorists at Balakot and shot down a Pakistani F-16. There was enough ambiguity around it that people who wanted to believe that narrative—a substantial portion of India’s electorate presumably—could. And these tools allowed the BJP to monopolize that narrative, particularly on WhatsApp. Pakistan was able to credibly claim the converse; the international open-source community concluded that India largely missed at Balakot and failed to kill an F-16, ratifying the Pakistan Army’s narrative. For Pakistan, Twitter worked in its favor, so it promoted tweets by “foreign experts” that validated Pakistan’s claims. Both sides were able to walk down from the crisis without further escalation despite it being the most kinetically aggressive engagement between two nuclear powers, perhaps in history. As Mark Bell and Nicholas Miller’s chapter in this volume suggests, it is unclear whether both governments opportunistically

used these fictions as cover to de-escalate or whether, more concerningly, they actually believed these fictitious narratives, which may incentivize future escalation. Nevertheless, this episode shows that social media platforms do not have a uniform effect on crisis dynamics and that platform properties, crisis properties, and audience properties interact in interesting ways to generate different dynamics.

HAWAII MISSILE ALERT 2018

On January 13, 2018, at 8:07 a.m. Hawaii time, every cell phone user in Hawaii received an alert through the official state government Emergency Alert System: “Emergency Alert: BALLISTIC MISSILE THREAT INBOUND TO HAWAII. SEEK IMMEDIATE SHELTER. THIS IS NOT A DRILL.” This was a push notification through the state’s cell networks, which essentially served as a universal text message to every cell phone on the Hawaiian Islands. At the same time, an emergency message scrolled on all local Hawaiian television stations: “If you are indoors, stay indoors. If you are outdoors, seek immediate shelter in a building. Remain indoors well away from windows. If you are driving, pull safely to the side of the road and seek shelter in a building or lay on the floor.” The alert was pushed with no context and no ability to otherwise determine whether the alert was genuine. It seemed authentic, and it was also plausible that North Korea had launched a missile at the United States, given recent events and barbs exchanged between Trump and Kim Jong Un. Many people took to social media to try to confirm the alert, but confusion ensued for what may have been excruciating minutes for the residents of Hawaii. This panic was exacerbated by the medium of text messaging, a platform akin to a social media tool: one with almost universal reach but on a closed platform with no ability to crowdsource accuracy such as on Twitter. It was not until thirty-eight minutes later, at 8:45 a.m. that a follow-up notification was sent to cell phones and TV emergency message scrolls that it was a false alarm.56

During those thirty-eight minutes, some social media platforms proved to be a more reliable source of information than official channels, and traditional news media largely relied on messages from Facebook and Twitter. Between 8:07 a.m. and 8:12 a.m., employees at the Hawaii Emergency Management Agency (Hi-EMA) notified state and local officials that it was a false alert and used the Hawaii Warning System to notify counties, “Attention all stations—This is the state warning point—There is no ballistic missile threat

Chapter 3

to Hawaii—this is a drill—I repeat, this is a drill."\(^{57}\) At 8:13 a.m., an employee at Hi-EMA posted that it was a “false alert” on their personal Facebook page, and five minutes later the agency’s PR team began sending notifications via social media. One of the first tweets confirming it was a false alert came from Rep. Tulsi Gabbard at 8:19 a.m.: “HAWAII—THIS IS A FALSE ALARM. THERE IS NO INCOMING MISSILE TO HAWAII. I HAVE CONFIRMED WITH OFFICIALS THERE IS NO INCOMING MISSILE.” This was confirmed one minute later by Hi-EMA and five minutes later by the governor. US Indo-Pacific Command apparently sent an email—with limited distribution by definition—stating it was a false alarm at 8:35 a.m., and at 9:08 a.m. it finally tweeted, “US Pacific Command has detected no ballistic missile threat to #Hawaii. Earlier message was sent in error and was a false alarm.”\(^{58}\) Traditional media sources and TV stations called on people to retweet their messages that it was a false alert in a classic example of cross-pollination between social media and news outlets.\(^{59}\)

Individuals and traditional news agencies were quick to respond, while the government response was relatively slow and careless. One reason for the delay between the two push notifications was that the Hi-EMA checklist for drills did not include a protocol response to false alerts. In order to issue a Civil Emergency Message (CEM) retracting the earlier alert, they needed confirmation of the code and requested additional guidance from the Federal Emergency Management Agency (FEMA). Hi-EMA first contacted FEMA for approval at 8:26 a.m., but could not get through until 8:30 a.m., at which time FEMA advised to issue a CEM. The CEM was drafted at 8:32 a.m. and pushed at 8:45 a.m.\(^{60}\) In an additional blunder, at 11:15 a.m. Hi-EMA tweeted that there would be a press conference at 1:00 p.m. about the “missile launch.” This was subsequently removed.\(^{61}\)

Both state and federal reports on the false alert blame “human error.” The original message was sent because an operator inadvertently selected “test message” instead of the “drill” option on a drop-down menu. Further, the


operator claims not to have heard, “EXERCISE EXERCISE EXERCISE,” and instead heard, “This is not a drill.” He had a record of poor performance, and according to a colleague, “He is unable to comprehend the situation at hand and has confused real life events and drills on at least two separate occasions.” After accidentally sending the missile alert message, the employee allegedly “just sat there” and did not “assist in the process” of canceling the alert. The employee later told NBC, “I was 100 percent sure that it was the right decision, that it was real.”

Although there was no other indication that Kim Jong Un’s North Korea was about to launch a first strike against Hawaii or the United States, this accidental alert came just six weeks after North Korea tested its third and longest range ICBM, the Hwasong-15. And it came only ten days after President Trump tweeted: “North Korean Leader Kim Jong Un just stated that the ‘Nuclear Button is on his desk at all times.’ Will someone from his depleted and food starved regime please inform him that I too have a Nuclear Button, but it is a much bigger & more powerful one than his, and my Button works!” For almost an entire year, the general public had been subjected to reports of almost biweekly missile tests (seventeen in total), a (likely) thermonuclear test in September, media reports painting Kim Jong Un as “crazy,” and presidential threats of “fire and fury.” The diplomatic charm offensive that Kim Jong Un would soon initiate had not yet begun. Furthermore, Hawaii was a ripe first target for North Korea due to its centrality for Pacific Command and its proximity to North Korea. It would have been plausible for an average citizen in Hawaii to believe that Kim Jong Un had decided to launch a first strike at the state.

In an added twist, it was roughly 2 p.m. on the East Coast of the United States when reports of the alert circulated on social media. President Trump was reportedly at Trump International Golf Course in Florida when those tweets began, though where he was precisely at this time is publicly unknown. At the time, Trump did not address the false alert. Gabbard tweeted, “Donald Trump is taking too long. Now is not the time for posturing. He must take this threat seriously and begin direct talks with North Korea, without preconditions, to de-escalate and denuclearize the Korean peninsula. There is no time to waste... The people of Hawai‘i should never have had to go through this.”

64. As discussed in “FEMA’s Oversight of the Integrated Public Alert & Warning System (IPAWS).”
particular delight in the “tragicomedy” and that “The entire island was thrown into an utter chaos at the news that a ballistic missile was coming in.”

But a hypothetical scenario helps demonstrate how dangerous this false alert could have been: what if President Trump had logged onto his personal Twitter account on the back nine and saw a screenshot posted by reputable users in Hawaii of the inbound missile alert and thought, “time to launch a counterforce strike against North Korea.” On the golf course, his chief of staff is nowhere to be found, but he summons his military aide with the nuclear satchel and orders to be patched through to the National Military Command Center (NMCC). He cracks open the “biscuit” that he keeps in his oversized golfing pants, validates and authenticates himself to the duty officer at the NMCC, and orders a first strike against North Korea, citing the inbound missile alert he has seen confirmed by the Hawaiian government officials. It is a legal and valid order, and no one could legally stop him.67

What does this incident reveal about social media during crisis? First, open social media platforms can provide an important source of information and quickly revert to the mean in disseminating accurate information. In this case, Twitter proved to be a better means of alerting the public than government officials or traditional news platforms. While the initial push message, essentially a closed platform, was alarming, subsequent open and closed platform messaging was highly effective in de-escalating tensions rather than stirring up fears of an incoming North Korean missile.

US FORCES KOREA NONESSENTIAL PERSONNEL EVACUATION ORDER 2017

Several months earlier, at the height of US-North Korean tensions, on the morning of September 21, 2017, a message was sent from the USFK Facebook account, on WhatsApp and via text message, calling for a noncombatant evacuation operation order (NEO). The order appeared to be official and from legitimate USFK and DoD accounts. It stated: “Real World Noncombatant evacuation operation order issued. All DoD family members and non-emergency essential DoD on the Korean Peninsula, an evacuation order has been issued” and was received from “USFK Official Alert.”68 The order requires family members and all noncombat essential personnel to evacuate in anticipation of a conflict and is considered to be the leading indicator


of a possibly imminent conflict—and a potential attack by the United States and South Korea against North Korea. It is one of the first steps the Pentagon takes in anticipation of conflict. The order was distributed from seemingly legitimate (compromised) accounts. There was no reason to believe it was not true.

Upon reports of the NEO order spreading virally within the USFK community, the latter quickly issued a denial on Facebook, which read:

On Thursday, 21 SEP, we received multiple reports of a fake text-to-cell and social media message regarding a “real world noncombatant evacuation operation (NEO) order issued” which instructed DoD family members and non-emergency essential DoD civilians on the Korean peninsula that an evacuation order had been issued. USFK did NOT issue this message. All US Department of Defense (DoD) family members are reminded to confirm any evacuation-related communications with their service member and unit non-combatant evacuation (NEO) representatives. Anyone receiving this false message should not click any links or open any attachments included in the correspondence.69

The closed nature of the platforms made it difficult to immediately counter the disinformation as it went viral with no widespread correction. Luckily, USFK acted swiftly and, interestingly, used the same platform—Facebook—on which the initial fake orders were posted, to correct the misinformation rather than sending an internal email or posting the correction on Twitter. An internal email may not reach as far as the original misinformation, failing to fully correct it, while tweeting a clarification may have reached too broad an audience, creating more panic than it solved by alerting those that did not see the initial Facebook message that there was a fake evacuation order even issued in the first place. Therefore, the sensible strategy was to use the same platform that propagated the disinformation in order to correct it. The DoD opened an investigation as to the source of the false message, and suspicion turned to a North Korean psychological operation, though as far as the authors are aware, there has been no official confirmation of this. If it was North Korean psyops, using legitimate and trusted, but compromised, social media accounts of USFK personnel added credibility and generated real concern that the evacuation order—and an imminent conflict—was genuine.

Given the intensity of tensions at the time, it would have been perfectly believable that the Trump administration had decided that military force was the only way to remove Kim Jong Un and his nuclear weapons. The order came at the height of the 2017 crisis between North Korea and the United States when it appeared that a war between the two was more than just a

remote possibility. Three weeks earlier, North Korea tested a purported thermonuclear weapon with a yield over 150kt.

Moreover, the United States conducted numerous exercises in the region immediately prior to and after the incident. In June and again in October 2017, it conducted genuine NEO drills, although officially the exercises are not in relation to any current events. A spokeswoman for the 2nd Combat Aviation Bridge told Stars and Stripes during the April exercise, “Obviously the one (scenario) that is forefront in our mind is a resumption of hostilities with North Korea or rising political tension leading to the feeling that American citizens might not be safe.” Earlier that year, in April 2017, Trump tweeted, “North Korea is looking for trouble. If China decides to help, that would be great. If not, we will solve the problem without them! U.S.A.” At the same time he told Fox News, “We’re sending an Armada” to include an aircraft carrier to deter North Korean aggression and further escalation. The announcement came only two days after a Trump-Xi summit in Mar-a-Lago, during which Trump expressed his frustration and waning patience with Kim Jong Un. The news about the carrier prompted a flurry of news stories, sparking fears of a US preemptive attack; however, the aircraft carrier, the USS Carl Vinson, was 3,500 miles away, traveling in the opposite direction. It was not until a week later, after completing a joint exercise in the Southern Pacific region, that the Carl Vinson began sailing north. During this time, Trump’s tweets and the clear confusion within the administration heightened fears of a possible escalation.

This episode shows how social media platforms can be misused to spread malicious and potentially war-hysterical information quickly, but also how that misinformation can be corrected relatively quickly as well. In this case, the crisis was embedded within a larger North Korea-US crisis, but the nature of the disinformation did not require action on the order of minutes—as in the Hawaii missile alert—since evacuations take days. Therefore, a considered denial on the same social media platform from the official USFK account sufficiently corrected the disinformation campaign. Twitter was used only once by USFK during the confusion, linking to the Facebook post clarifying that the message was fake. Similarly, the Navy and other Services retweeted the USFK message or news stories from Stars and Stripes and Military Times about the fake order. Again, this points to a cross-pollination

relationship between social media and more traditional media sources, such as *Stars and Stripes*. Whereas in the Hawaii false alert case, social media was out in front of traditional media and played an important role in public communications, military news outlets were the main source of information for the USFK case, aside from the Facebook post by USFK. This case is striking in the lack of social media activity it generated, and the *Stars and Stripes* story dominated Twitter activity about the case, accounting for the top three most frequently tweeted/retweeted posts about it in the immediate aftermath, aside from the USFK original tweet and link to the Facebook story.74

**Conclusion**

This chapter has attempted to illustrate that social media and its impact on crises between nuclear actors is complex and not a monolithic phenomenon. We make the simple but powerful point that properties of the platform, the crisis, and the intended audience can interact in interesting ways to lead to different types of dynamics in crises between nuclear powers, some escalatory and some de-escalatory. They do not have a uniform effect. The analysis of the three cases demonstrates the open and closed social media platforms were used differently and to different effects. This provides strong support for our first hypothesis, that social media does not have a uniform effect on crisis dynamics but can have varying impact depending on the specific type of platform and the nature of the crisis.

Did open social media platforms do more to thicken the fog of war than closed platforms during discrete crises? Not exactly. Open platforms, such as Twitter, during both acute and ongoing crises, invite more information sharing so as to ascertain a holistic picture closer to the truth. This, too, is reinforced by studies into the use of Twitter, with one study finding that “far from safely agreeing with the dominant opinions in their respective social communities or merely being expressive, users debated with and challenged one another, even inviting disagreement.”75 And did closed social media platforms reinforce existing views over ongoing crises to include conspiracy theories? Yes, at least in the Pulwama/Balakot case.

It was more challenging to prove or disprove our second hypothesis that social media does not have a uniform effect on international and domestic audiences. We have hypothesized directions and effects of these dynamics,

74. One of the other most retweeted stories was by a prominent Japanese academic confirming that the order was fake. Okuyama Masashi, Twitter, September 21, 2017, https://twitter.com/masatheman/status/9110013231564802.

but do not have enough data—some of which is incredibly difficult to observe and collect—to conclusively answer these questions, which are critical to future nuclear-tinged crises. In addition to these questions, the research also raises important questions about wider nuclear strategy in two specific ways. First, it highlights risks and important questions about chain of command and leadership during crises. In the Hawaii case, for example, the first person to publicly announce that the alert was fake was one of the employees at Hi-EMA who posted to his personal Facebook account. And the USFK example suggests that social media can be manipulated potentially by malicious actors. This begs the question: if orders can be issued via social media, will they be subject to similar vulnerabilities? The Twitter hack on July 15, 2020, in which several hundred high-profile “verified” accounts, including the then-presidential candidate Joe Biden’s, were taken over in a bitcoin scam—perpetrated with insider help by a Florida teenager—raises terrifying prospects. What if that hack were executed at the height of a military crisis and the perpetrator did not want bitcoin but to start a war? Imagine if President Trump’s account had been taken over and a hacker simply tweeted “Game over Iran” or “Your time is up Kim.” North Korea appeared to monitor President Trump’s Twitter account in real time and having little early warning capability, may be forced to have an itchy trigger finger if it fears an imminent US attack. The power of some of these platforms, and their inherent insecurity, can generate some hair-raising scenarios.

If social media can contribute to nuclear escalation, should governments close down social media during a crisis? There are no easy answers to this question as a shutdown may create more rumors or risks—and remove potential pathways to de-escalation—than keeping the platform fully operational. In late 2019, for example, Iran shut down the country’s access to the internet, including both Iranian and international sites, for a week amid ongoing protests over rising fuel prices. Iran, Russia, and potentially other countries are refitting their national internets to make it easier to shutdown access at the government’s behest, similar to China’s “Great Firewall.”

According to internet freedom groups, outages are becoming increasingly common and jumped from 75 to 196 between 2016 and 2017. Social media and its cross-pollinating effects will remain an important force that governments will have to contend with in future nuclear-tinged crises.

Social media is rich in variance. Different platforms—by design—make it harder to control the narrative, whether to deploy nationalist fictions or withhold inconvenient facts. Some platforms are dangerous vectors for dis-


information, as suggested by two of the incidents offered here, which were false alerts. This may change the ways leaders communicate the crisis and spill over to public or use public domain for communicating, making private diplomacy ever more challenging and complicated. For example, Vladimir Putin’s order to increase Russia’s nuclear alert level during the 2022 Ukraine war went viral on Twitter, generating anxiety about a nuclear crisis between the United States and Russia. Yet it was also nuclear experts on Twitter that reduced global anxiety by explaining that Putin’s order was largely rhetorical, with no practical impact on Russia’s force posture. In this case Twitter served as a platform that both increased and subsequently reduced public anxiety about the nuclear crisis. Our fundamental conclusion is that different social media platforms have varied volume, accuracy, and speed of information leading to distinct effects and potential pathways to escalation and de-escalation in international crises. The only uniform effect that social media as a whole injects into nuclear crises is complexity. And over time, complexity in more frequent and potentially intense nuclear crises can generate a bill no one wants due.