Toda Peace Institute

Policy Brief No. 66 **November 2019**

Social Media Arrives on the Nuclear Stage[1]

Peter Hayes

Introduction

Social media burst onto the stage of nuclear warfare in 2018. In the Asia-Pacific region alone, six instances of social media playing a role in nuclear-prone conflicts occurred between August 2017 and January 2018. Three (September, November, December 2017) related to indicators that the United States might be readying to attack North Korea with nuclear weapons.

---

[1] This brief draws from Nautilus Institute, Technology for Global Security, Preventive Defense Project, "SOCIAL MEDIA STORMS AND NUCLEAR EARLY WARNING SYSTEMS: A DEEP DIVE AND SPEED SCENARIOS WORKSHOP REPORT", *NAPSNet Special Reports*, January 08, 2019, https://nautilus.org/napsnet/napsnet-special-reports/social-media-storms-and-nuclear-early-warning-systems-a-deep-dive-and-speed-scenarios-workshop-report/ under a 4.0 International Creative Commons License and was funded by the MacArthur Foundation.
At the workshop, analytical materials were reviewed and then a speed scenarios exercise involving nuclear weapons and social media experts and practitioners was conducted to explore antidotes to potentially catastrophic effects of social media on the risk of nuclear war. Additional expert papers were published from this workshop and are drawn upon in this brief.
Renée DiResta, "OF VIRALITY AND VIRUSES: THE ANTI-VACCINE MOVEMENT AND SOCIAL MEDIA", NAPSNet Special Reports, November 08, 2018, https://nautilus.org/napsnet/napsnet-special-reports/of-virality-and-viruses-the-anti-vaccine-movement-and-social-media/
Brittan Heller, "THE ONLINE TARGETING OF JOURNALISTS WITH ANTI-SEMITIC INTIMIDATION", NAPSNet Special Reports, October 25, 2018, https://nautilus.org/napsnet/napsnet-special-reports/the-online-targeting-of-journalists-with-anti-semitic-intimidation/
Sunil Dubey, "CITIES, SOCIAL MEDIA, AND PREPAREDNESS FOR MAJOR THREATS", NAPSNet Special Reports, November 15, 2018, https://nautilus.org/napsnet/napsnet-special-reports/cities-social-media-and-preparedness-for-major-threats/

TABLE 1: 20170-18 FALSE ALARMS AND SOCIAL MEDIA STORMS ABOUT NUCLEAR AND MISSILE ATTACKS

GUAM: E. Adamczyk, "Guam residents unnerved by accidental 'civil danger warning,' UPI, August 15, 2017.

SEOUL: D. Lamothe, "U.S. families got fake orders to leave South Korea. Now counterintelligence is involved," *Washington Post*, September 22, 2017.

"The U.S.S. Kentucky is part of what is called the nuclear triad." The triad are the three components of a nuclear defense system: "land-based missiles fired from secret silos, B-1 bombers that can drop them from the air, and submarine-launched ballistic missiles." STRATCOM tweet, C. Perez, "Military tweets out error-filled story about US nukes," *New York Post*, November 15, 2017.

SOUTHERN CA: 'Update: Minuteman III ICBM test launch from Vandenberg canceled," December 6, 2017 Postponed" Minuteman missile launch supplanted by unannounced Trident missile launch same day
(April 7 2013 launch postponed to avoid provoking DPRK.)

HAWAII: A. Wang, B. Lyte, 'Ballistic Missile Threat Inbound To Hawaii,' the alert screamed. It was a false alarm," *Washington Post,* January 13, 2018 .

TOKYO: AP, "Japan public TV sends mistaken North Korean missile alert," January 16, 2018.

Three (August 2017 and two in January 2018) led to social media storms that amplified false alarms of pending nuclear attack involving millions of people hiding under tables and waiting for their world to end.

For decades, strategists have worried about the possibility that states armed with nuclear weapons might mistakenly launch a nuclear strike due to a false alarm originating in its early warning system or due to degraded decision-making. (The flip side of that concern is that a nuclear weapons state might not notice that it is under nuclear attack because of errors in its early warning system and might not respond appropriately due to degraded nuclear decision-making).

Nine states now have nuclear weapons. Each of them has a nuclear command and control system to maintain their nuclear forces, and to operate them in peacetime. Those who command these forces rely on information about the status of their potential nuclear adversaries. This information is obtained from many sources culled together in strategic intelligence; and on sensors and other sources of real-time or immediately available information that monitor the status of potential attacking that, considered together, suggest that nuclear weapons are—or are not—about to attack a given nuclear weapons state.

Thus, each nuclear weapons state maintains an early warning system that evaluates threat data, and if it receives information that might suggest an attack is underway, assesses the significance of the threat. Some states use physical sensors such as infrared detectors on satellites and long-range radars to provide "dual" warning from physically distinct and separate systems; and rely on one to cross-check and confirm the readings from the other. In this regard, the United States has the most mature early warning and nuclear decision-making system, supported by a global network of sensors linked by communication systems to early warning assessment centres that in turn report to nuclear commanders. (See Table 1).

Other nuclear weapons states have much less capable early warning systems, using a few satellites with limited coverage, supplemented by a few other long-distance sensors such as radars. Some have no long-range sensor systems at all, such as North Korea.

Yet all these states have social media available to their officials, even in North Korea. Thus, the first warning that nuclear capable missiles or bombers might be heading for Pyongyang might be on Twitter or Facebook. In addition to providing possible early warning of the physical status of nuclear weapons, social media may also provide unprecedented and unique access to the intentions—and the state of mind—of a nuclear commander.

Even more worrisome, much of the information found on social media is factually incorrect; and some of it is purposely posted to manipulate users as "fake media" or to fan the flames of conflict by manipulating readers *en masse*. Even worse, it is becoming difficult and even impossible to distinguish between actual videos, photographs and voices, and "deep fakes."

Why does this matter? The primary reason is the terrifying combination of speed and the unique scale of violence when it comes to nuclear weapons that continues to set them apart from all other means of coercion. The simple problem is that nuclear commanders must make decisions to use nuclear weapons for mass destruction in time measured in minutes and seconds, not hours, days and minutes. This is due to the compression of decision-making time by the deployment of long-range delivery systems that can take as little as 10-12 minutes to arrive from firing point.

How should a nuclear weapons state treat the vast amounts of social media, including content that may be factually accurate that is transmitted almost instantly, but may also be created by malevolent parties that aim to deceive, manipulate, and mislead its early warning system and pollute its nuclear command decision-making process with fake information?

In addition to "traditional" sources of NC3 error such as accidents, hardware failure, and human error, new technologies such as artificial intelligence, autonomous vehicles, quantum computing and sensing superimposed on the legacy US NC3 system may create new types of coincident error involving social media.

In the past, the insulation of nuclear commanders from unclassified data and their near total dependence on official, classified information systems in the midst of crisis might have served to reduce the influence of erroneous or deceptive information. In today's world where the President of the United States makes public declaratory policy on Twitter, one cannot be sure that social media will not be influential.

Recognising the significance that nuclear commanders now inhabit a social media universe, in this early warning-assessment-alerting-decision making sequence, social media posts are only one of many inputs to interpretation. Social media may provide supplementary data or influence that—when combined with other sensor data and indicator— "tip" the assessment from "no attack underway" to "possible attack" or "attack underway."

In our study, we posit that social media's influence is always at the margin, supplementing information from strategic intelligence (such as communications or electronic intelligence)

and real-time sensor data, or tactical warning; and that no nuclear commander would ever make a decision solely based on social media posts by an adversary, nor by a friendly social media source that is valued by a nuclear commander.

Lessons from Social Media in Non-Nuclear Domains

To anticipate how social media might play out in the world of nuclear early warning, we considered studies of social media in other domains where it was used to promote extremist views and behaviour in promoting anti-vaccination,[2] anti-Semitism,[3] gang, ethnic, and terrorist violence in cities.[4] Lessons learned from these studies are summarised below.

The first set of lessons concerned the question of false alarms generated by social media, and the resultant creation and amplification of conflict where little or none existed before.

In the case of the "anti-vaccers" or social media campaigners who aim to stop vaccination, a case study showed that a virtual social network is vulnerable to cross-platform manipulation that develops a large standing audience for the anti-vaccination perspective. The anti-vaccers used many sophisticated techniques to drive vulnerable readers away from vaccination such as the use of gameable algorithms. The published paper by Renee Diresta argues that in the anti-vaccination case, a confluence of three factors - mass consolidation of audiences onto a handful of social networks; the adoption of curatorial algorithms as a primary means of disseminating and engaging with content; and the ease of precision targeting of users via the leveraging of proprietary profiles built from their own media consumption signals - has resulted in an information ecosystem that can be manipulated by a variety of actors with relative ease.

Concludes Diresta:

> The anti-vaccine movement is well-funded and technically savvy. They followed the best practices of internet marketers, writing blogs and cross-promoting content and sharing material across all of the new platforms. Social network design choices meant that popularity determined what people saw; even nuanced policy issues began to be run as digital marketing campaigns.
>
> In effect, the anti-vaccers used social media on a massive scale to short circuit the traditional flows of medical knowledge and expert, evidence-based advice available to individuals (in particular, parents) to reduce the rate of vaccination, in some cases, to the point that public health was threatened by revived outbreaks of contagious diseases.

---

[2] Renée DiResta, "OF VIRALITY AND VIRUSES: THE ANTI-VACCINE MOVEMENT AND SOCIAL MEDIA", NAPSNet Special Reports, November 08, 2018, https://nautilus.org/napsnet/napsnet-special-reports/of-virality-and-viruses-the-anti-vaccine-movement-and-social-media/

[3] Brittan Heller, "THE ONLINE TARGETING OF JOURNALISTS WITH ANTI-SEMITIC INTIMIDATION", NAPSNet Special Reports, October 25, 2018, https://nautilus.org/napsnet/napsnet-special-reports/the-online-targeting-of-journalists-with-anti-semitic-intimidation/

[4] Sunil Dubey, "CITIES, SOCIAL MEDIA, AND PREPAREDNESS FOR MAJOR THREATS", NAPSNet Special Reports, November 15, 2018, https://nautilus.org/napsnet/napsnet-special-reports/cities-social-media-and-preparedness-for-major-threats/

In the case of anti-Semitic online activism, a case study by Brittan Heller found that the trolling and attacks on minority journalists, especially of Jewish ethnicity, used social media-based aggression to spark violence, including off-line violence. This study found that the online attacks were highly targeted—83 percent of 2.6 million anti-Semitic tweets, for example, were targeted at only 10 people; and that the attackers were professionalised, not "amateurs." Moreover, stated Heller:

> Overall, the study found that a comparatively small group of attackers drove most of the anti-Semitic hate and harassment on Twitter, but these individuals had an outsized impact. More than two-thirds of the anti-Semitic tweets directed at journalists were sent by 1,631 Twitter accounts, out of 313 million total Twitter accounts at the time of the attack. While this is a small proportion of Twitter users, the comparative impact of this abuse was widespread. The reach was tantamount to the spread covered by a $20 million-dollar Superbowl ad.

As with the anti-vaccers, a relatively tiny number of social media activists were able to manipulate readers to change their views and their behaviour in the real world, by not vaccinating or by attacking target persons not only virtually but in the real world. Other examples from other domains were cited at the workshop along the same lines.

In cities, social media has been employed to motivate and then to orchestrate mob violence against minority populations, and even to coordinate large-scale terrorist attacks (as in the case of Mumbai in 2008.) Cities are natural targets for such manipulative campaigns because they aggregate so many people who can be networked into flash mobs and riotous behaviour for which traditional policing has no answer short of total shutdown of civilian communications such as occurred in London in 2012. Thus, Sunil Dubey concluded that:

> By 2030, over 65% of total world population will live in cities. Cities confront the rising influence and penetration of social media platforms on all aspects of urban life. Although this virtual urban life makes cities smarter, more efficient, and more sustainable in many respects, it also subverts the safety, security and resilience of our cities."

In cities such as Chicago, as much as 80 percent of gang violence starts or is facilitated online. Thus, time and again, social media is proving capable of quickly mobilising fear and channeling that aggregated animosity against real-world targets. In these domains, therefore, there are precursors of how activists might attempt to mobilise mass online campaigns to generate fear and alarm in the face of nuclear threat; and to target key individuals either in a nuclear command and control system, or nuclear commanders themselves, to launch a nuclear attack against a third party presented as worthy only of nuclear annihilation. Such campaigns might also involve the use of fake news propagated over social media to justify the campaign—as occurred on September 11, 2104 in an online hoax involving many fake twitter accounts that posed about the attack and targeted celebrities in order to maximise the attention. Not only did the hoaxers present edited CNN screenshots; they even posted functioning clones of TV stations that purported to cover the event.

This creation of a virtual attack is similar in nature to the fake post of a non-existent non-combatant evacuation in Korea in September 2017. It suggests that it is almost inevitable that individuals, organisations, and even states may start to use social media to try to provoke nuclear attacks against their adversaries; or for other political-ideological or religious reasons; that they will be effective in terms of reaching some highly influential people as well as large numbers of people; and included in these two types of readers are likely to be some people making nuclear early warning assessments, and nuclear command decisions.

The second set of lessons concerned the question of what antidotes exist for these types of false alarms either on social media, or via other ways of creating authoritative and credible reference knowledge. A variety of strategies were described that partly ameliorated the problem, although generally the response by social media platforms themselves was slow, often only at the behest of outsiders, and inhibited by many problems of attribution and tracking of online aggressors.

In one real world circumstance involving crisis management with North Korea, it emerged that social media plays out differently in the Korean context than in the west (there being relatively fewer Twitter users in South Korea). In general, social media widens the pre-existing political and ideological divide that characterises almost all public discourse in South Korea, and after a land mine exploded in 2015 injuring South Korean soldiers, social media sought to punish the North. Even the fake non-combatant social media report was quickly dampened by countervailing messaging by US Forces Korea.

Where social media platforms have attempted to curtail manipulative and dangerous use of their services, they have found that one of the best ways to do so is to simply slow down the ability of users to run their campaigns. This can be achieved by automated systems that shut down sites found to be fake sites and by identifying core behaviours used by social media activists that become markers that can be used to create traction and slow down their ability to game the rules of social media by using bots, anonymity, etc. One problem in such regulation of online behaviour is that context is critical to determining what is and is not dangerous, and a human must be in the loop. However, once deep fakes become widespread, even having a human involved may not suffice to determine the truth content of a specific post or site.

Given the speed of nuclear early warning and nuclear command decisions, following these pathways is likely too little, too late. Indeed, it is unclear who social media practitioners should tell if and when they find something portentous of nuclear attack on their platforms.

Thus, a "truth infrastructure" trusted by users is required to provide legitimate and authoritative review of what's real versus what is fake. As the number of points of governance in the nuclear weapons field increases—in part due to the proliferation of the weapons, and in part due to the involvement of more actors lower in the governance system, especially of cities—it seems clear that a pre-existing source of authoritative information on the status of nuclear forces that is judged to be credible by nuclear weapons states and independent

of their own and their adversary's early warning systems is the only sure-fire way of overcoming the pernicious effect of social media on early warning assessment and command decisions.

As nuclear command and control is defined by the imperative to make decisions in the context of inevitable uncertainty, what becomes important in a world saturated with social media posting is not just finding an answer—of which there are any number of competing claims—but having to search for the reliable answer given that you get faster, instant answers on social media, and many people accept the first answer that comes up.

One possibility to reduce uncertainty is that first responders, especially at the city level, might use their information systems to provide credible information to reporters and mass media; and to fold carefully evaluated social media reports into their information after first validating the early reports with a variety of real-time sensors. First responders noted that it is essential to ensure that the messages across cities and counties are aligned and consistent, to avoid public confusion and disenchantment with official sources of early warning. However, first responders know from experience that a whole battery of circuit breakers to preempt and steer false information away is needed, not just one message.

One case in point is that of Hala Systems which provides early warning of pending bomb attacks to civilians trapped in rebel-held areas of Syria.[5] Hala observe aircraft using a variety of data-mining techniques to predict attacks and sends Take Cover alerts via Telegram and Facebook platforms. In the case of ShakeAlert,[6] an earthquake early warning social media App that uses smart phones to collect seismic data by detecting ground motion, and then collates and interprets the data to predict location and intensity of the pending earthquake sufficiently quickly to allow critical infrastructure and individuals to take immediate protective steps (that is, within 15 seconds), the system relies solely on physical sensors and automated interpretation and communication.

Thus, even in the case of threats with extremely short timelines—minutes in the case of ombing, seconds in the case of earthquakes, it is possible to use social media to inform and guide humans to respond in ways that reduce risk, without the risk of the alert system being hijacked by malevolent actors. The critical element is that the users trust the reliability and authenticity of the information sent out—just as it would be for an independent, impartial third party nuclear early warning system.

These considerations led to posing of key questions for exploration in the speed scenarios.

For nuclear armed states:

- Should nuclear early warning systems include social media in their threat assessments?
- Should nuclear early warning systems ignore social media reports of attacks to avoid increasing frequency of assessment and of eventual assessment error?

---

[5] https://halasystems.com/
[6] https://www.shakealert.org/

- Should they rely on the maturity, competence and professionalism of their adversaries to assess the status of their own nuclear forces, or is it time to start building collaborative information systems that reduce the risk that they may make errors, including errors that might arise from social media reports?

For everyone, including non-nuclear states, and non-state sectors such as social media:

- Is there a third party that can provide real-time status of nuclear forces that would serve as an independent reference for nuclear weapon states early warning systems and commanders and everyone else?
- If so, who should take the lead in creating it?

These and other questions formed the basis for the Speed Scenarios that are described in the next section of this report.

Circuit Breakers and Short Circuits Speed Scenarios

Four "short circuit" hypothetical, imaginary scenarios were produced at the workshop that explored how and what circuit breakers might be created that avoid or overcome the destabilising effect of social media on nuclear early warning systems and nuclear command decisions.

The four scenarios and circuit breakers include:

Korea A Sweltering Crisis—The United States and North:

Remember that kumbaya moment in December 2018, when North Korea declared it would dismantle most of its key warheads and missile facilities? All that ceremonial press coverage showing the North Koreans handing over fissile material to China and destroying warhead parts under the supervision of joint US-China-Russia teams working with the IAEA, all of it complete by June 2019? Oops. Guess who still has nuclear weapons? Within minutes of the North Korean aircraft plummeting to ground, Kim Jong Un defiantly announces that he will deploy ten previously undisclosed, nuclear-armed, long-range missile launchers. The message to the United States and South Korea is crystal clear: Stand down, or we will unleash the largest nuclear strike in history.

In this hypothetical scenario, only summarised here, the United States is already on red alert, a fact communicated that evening by a Presidential Alert message that hits every American mobile phone. So are the South Koreans, and so is Japan. Having veered sharply from Article 9 of its Constitution (which prohibits waging war) toward a much more assertive posture throughout East Asia, in this imaginary scenario Japan now has at least 20 nuclear weapons of its very own, although none have been tested except on supercomputers.

As the world contemplates disaster, tempers explode, and temperatures soar, millions of Twitler accounts suddenly start lighting up with the same seemingly official message, sent and re-Twitled so quickly that it's hard to figure out its origins: *All US and allied non-combatants on the Korean Peninsula are to report for immediate evacuation.*

Multiple circuit breakers were prefigured including launching fact-finding missions, improving official communications with mass media, and leveraging trusted third parties willing to put themselves on the line as exchange-hostages to push the international community to a peaceful solution, reestablish trust, and mitigate panic.

Fast and Furious—India, Pakistan, and a Non-State Actor :

In this hypothetical scenario it is December 2021, and we are in the midst of a news cycle like nothing we have ever seen in the India-Pakistan nuclear standoff. After five days of rapid mutual escalation, a newly established terrorist group detonates an unknown weapon in the hijacked city of Kalpakkam, with widespread damage at the atomic reactor facilities. While details are unclear, it appears that a commandeered LNG tanker was used and the explosion was immense. The role and likely spread of radiological materials also remain unclear, but the attack is live-tweeted by terrorists, victims, and global intelligence services, effectively turning the catastrophe into a global media event. Pakistani social media report that a Pakistani nuclear-armed submarine is preparing to put to sea, based on posts by its crew to their families. India puts its air force and missile force on high alert, begins massing significant ground forces, and puts Pakistan on warning. Pakistan reciprocates by putting its military forces on their highest level of warning.

On the sixth day of 24/7 crisis, the Pakistani prime minister responds with a Tweet that threatens India and others with nuclear reprisal if any military action is taken. Without mentioning the United States by name, he also asserts that any attempt to intervene with special forces to attack Pakistan-based terrorists will be met with an "instant and devastating response". He calls on all Pakistanis to monitor the skies day and night, and to use social media to instantly report any incursions by "outsiders."

No one knows where this set of spiraling threats and attacks is heading. What is clear is that non-state actors have skillfully executed nuclear threats and attacks in ways that have caused multiple nuclear-armed states to threaten one another—and that their plot has launched nuclear escalation spirals across the global landscape.

The circuit breaker in this firestorm scenario aimed to create the necessary political space to deescalate the situation, starting with mitigation strategies set in motion by social media luminaries and companies to slow down communication combined with inter-state back-channel conversations and more traditional mutual hostage taking to stop conflict from spiraling out of control in conditions of extreme nuclear provocation.

Mutual Miscalculations—NATO, United States, and the Russian Federation:

Seizing an opportunity created by most nations' preoccupation with domestic affairs, this hypothetical scenario involved Russian forces moved into Belarus, escalating tensions with NATO. Concern about Russia's intent on its western border with Europe intensified. When Russia mobilised a large flotilla off the coast of Syria the distrust was amplified even more. And Vostok-2018 in the Pacific, involving 300,000 troops and 900 tanks—the largest military drill since the Cold War—with China also participating, significantly increased the West's concerns about the warming relationship (and economic collaboration) between

these two great powers, leading the US to strengthen ties with Poland and to speed up deployment of missile defenses in Poland.

All over the world, citizens posted impossible demands on their leaders: Avoid these wars! Protect us from attack! Up with NATO! Down with NATO! Up with Putin! Down with Putin! Up with Avenatti! Down with Avenatti! The clamour was both confusing and impossible to ignore. It also made it impossible to separate important signals from the noise.

When the US increased its security alert to DEFCON 3 with preparations for DEFCON 2 in place, Russian counterparts did the same.

And then ISIS made its move, exploding a small nuclear device in Syria but made no claim of responsibility, and all hell broke loose in the Middle East. Blame was placed on the US, Russia, terrorists. The violent accusations and denials were everywhere. There was confusion as to the whereabouts of all of the approximately 50 US nuclear weapons at Incirlik Air Base. No one could convincingly deny guilt. And now the US early warning system is showing incoming missiles.

Flanked by his closest advisors in Washington, US President Avenatti is focused on trying to make sense of conflicting information coming from American nuclear early warning systems. Nothing in his training as a litigator or political operative provided what he needs now: a way out of a dark and extremely perilous box. Putin has not picked up the legacy analog phone of the US-Russian hotline throughout the escalating crisis.

Avenatti is a true Twitter aficionado. He feels his Twitter finger twitch and reaches for his phone.

The circuit breaker set out to answer the question: Are there ICBMs in the air and if so from where? while dealing with increasing panic. The circuit breaker is to create a new international organisation with only one mission—to provide sensor-based information on international incidents that may related to nuclear war, to validate data events; and impartially to send data to all parties to a nuclear prone conflict.

Embrace Tiger, Retreat to Mainland—China, Taiwan, and the United States:

In this imaginary scenario, the newly elected Taiwanese government's declaration of independence from China sent political and military shockwaves across East Asia. Since the upgrade of its forces in the first decade of the 21st century, China's Second Artillery Force viewed itself as "the arrow on the bow and poised to strike," able to exert tremendous pressure on the "Taiwan separatists." Thus, Taiwan's declaration was a direct challenge to the PLA's core identity. It hankered for orders from the Central Military Commission to show Taiwan it could not ignore the military power of the mainland.

It took less than a day for Chinese President Xi Jinping to direct the Chinese Navy to impose a naval and air blockade around Taiwan. In return, Taiwanese forces fired a barrage of anti-ship missiles at Chinese forces, simultaneously launching long-range missiles at offshore islands where Chinese ground and amphibious forces are massing to invade Taiwan. The

United States Navy sent destroyers into the Taiwan Straits, and US strategic submarines departed Guam and the US West Coast for open ocean. A series of shootouts take place involving bombers and submarines. At the same time, a storm of social media attacks on different factions inside and outside of China takes place, with claim and counterclaim as to who is in charge, and whether China is firing missiles at the United States.

Now, in this hypothetical scenario, President Pence takes two minutes to kneel and pray. Then he stands up, ignores the Secretary of Defense, and requests an aide to assist with selection of nuclear strike options from the football—one for North Korea and one for China. He is advised that the only way to deliver nuclear warheads on these two countries is to fire them from the US Mid-West using land-based missiles, and that these will have to fly over Russia on their way to their targets, albeit in space, not Russian airspace. He shrugs.

Pence receives and notes a report from the CIA that there are no social media reports of these missiles taking off over the heads of millions of Chinese, all armed with smartphones, yet the contagious propaganda attacks continue. The CIA is unable to determine if crowd reporting of the launches is missing due to censorship controls, or if its absence is confirmation that in fact no missiles were launched and that the satellite early warning sensors have mistaken missiles for some other infrared signature. They recount a blitz of social media in China issued by the government and individual celebrities that China will pay any price to stop the United States from separating Taiwan province from the mainland and calling on the Chinese diaspora to rise up and strike against the United States everywhere in the world.

He asks how long before the missiles strike and whether it is certain that they will hit land or the ocean? He picks up his phone when...

Like the third scenario, the circuit breaker in this instance tries to reduce uncertainty at the brink: What is the nature of the Chinese missiles, nuclear or conventional, that are in the air; and will they hit Guam and Okinawa or splash down nearby? In this case, China initiates a last minute, last second concerted, all-out *diplomatic* effort to stop a US retaliatory strike by calling on trusted persons to act as intermediaries. They activate personal backchannels including businessmen, politicians, and even religious leaders. Their first act (because it is the fastest) is to turn off aggressive social media in China itself.

Conclusions

Some themes recurred across all the scenarios, suggesting that these elements might lend robustness to many de-escalatory strategies. For example, whatever the role of social media in creating or amplifying nuclear-prone conflicts, high-level communication between trusted individuals was an ingredient in resolving conflict. Other elements included high and low-level hostage exchange; doing whatever it took to slow escalatory spirals; and anticipating the loss of control induced by social media and other drivers by establishing hot lines, market and civil society-based communication channels, and trusted, third party, and impartial sources of authoritative information on the status of forces.

Thus, social media platforms and social media users can shift the center of gravity away from the current, celebrity-driven and social media conflict amplifying dynamic that degrades the quality of much information and towards more reliable, authenticated information while preserving the ability of users to free speech and near-instantaneous networking.

In this regard, cities and civil society emerged as a set of actors and networks that may be positioned to create new forms of governance and public information goods that restrain the aggressive use of social media that may contribute to false alarms and poor decision-making at the national level, while contributing to independent, impartial and validated information that is useful to nuclear early warning systems and nuclear commanders who may be relatively poorly served by traditional sensors, early warning systems, and conflict resolution mechanisms at the level of inter-state conflict. There are a number of possible "champion cities" already greatly concerned about their vulnerability to the effects of nuclear war, most notably some of the lead cities in the Mayors for Peace global network.[7]

However, there is much potential to expand this role to other global city networks such as Metropolis,[8] to integrate city-level nuclear war monitoring of the status of deployed nuclear forces and of the location and "ideational" and "emotional" status of civilian and military nuclear commanders in each nuclear weapons state.[9] Such social surveillance is already in place in areas in which genocide and other crimes against humanity exist. Training, equipping, and certification of such city level and "civil society" reporters might well contribute to nuclear risk reduction, and may be of interest to nuclear weapons, nuclear umbrella, and nuclear prohibition states, putting aside their other reasons to differ.[10] Such monitoring and reporting might be integrated with a generic disaster-driven risk management monitoring and reporting system rather than be nuclear-only, thereby maximising the possible set of cities with an interest in participating in a global network of civilian reporters. City-level monitoring and reporting could also be integrated with information provided by non-nuclear weapons states, whether they be nuclear umbrella states, nuclear prohibition states, or nuclear bystander states, using national technical means that might be beyond the technical ability of non-state civil society, local government, and other urban actors. When the ability of disruptive market entities such as cube-sat communications and remote sensing companies are added to the mix of possible suppliers of sensed data, a truly independent and reliable monitoring system is possible, given political will.

---

[7] See http://www.mayorsforpeace.org/english/
[8] See https://www.metropolis.org/
[9] For a trial run at such an index, this one on the "tone" or emotional status of state leaders, based on all media across all major languages, updated regularly, see See "New GDELT World Leaders Index," February 9, 2014, at: https://blog.gdeltproject.org/new-gdelt-world-leaders-index/
[10] See comments on distributed societal monitoring and surveillance by NATO Deputy Secretary General Rose Gottemoeller, in "NATO Nuclear Policy in a Post INF World," University of Oslo, September 9, 2019, at: https://www.nato.int/cps/en/natohq/opinions_168602.htm?selectedLocale=en

The Author

Peter Hayes is Honorary Professor, Center for International Security Studies, Sydney University, Australia and Director, Nautilus Institute in Berkeley, California. He works at the nexus of security, environment and energy policy problems. Best known for innovative cooperative engagement strategies in North Korea, he has developed techniques at Nautilus Institute for seeking near-term solutions to global security and sustainability problems and applied them in East Asia, Australia, and South Asia. His current project is to increase the accountability of nuclear weapons commanders. He advises the Asia Pacific Leadership Network and is on the editorial board of Global Asia.

Toda Peace Institute

The **Toda Peace Institute** is an independent, nonpartisan institute committed to advancing a more just and peaceful world through policy-oriented peace research and practice. The Institute commissions evidence-based research, convenes multi-track and multi-disciplinary problem-solving workshops and seminars, and promotes dialogue across ethnic, cultural, religious and political divides. It catalyses practical, policy-oriented conversations between theoretical experts, practitioners, policymakers and civil society leaders in order to discern innovative and creative solutions to the major problems confronting the world in the twenty-first century (see www.toda.org for more information).

Contact Us
Toda Peace Institute
Samon Eleven Bldg. 5th Floor
3-1 Samon-cho, Shinjuku-ku, Tokyo 160-0017, Japan
Email: contact@toda.org