A Potent-but-Patient Approach to Cyber-Deterrence

U.S. strategy has not kept pace with the evolving cyber threat. Recent proposals ignore key strategic features of the cyber domain, resulting in overly narrow policies. We must take a global approach to cyber-deterrence, and we must blend aggressive retaliation when the origins of attacks are clear with forbearance when they aren't.

By Sandeep Baliga, Ethan Bueno de Mesquita, and Alexander Wolitzky

Here's a scenario that should trouble America's political leaders:

Top-secret plans for a next-generation fighter jet are stolen from a U.S. defense contractor's computers.

It appears the intrusion originated in China. Then again, it's easy for other actors to make it look as if the culprit is China. And some signs point to North Korea.

Ultimately, the U.S. blames China. It launches a retaliatory cyber-strike that paralyzes Chinese military computer networks for a week. U.S. diplomats tell their counterparts they've been warned against future incursions.

But the move backfires.

It turns out the initial attack came from Iranian Revolutionary Guard operatives routing the attack through a server in China and using North Korean coding techniques.

Chinese leaders are incensed—viewing claims about misattribution of an Iranian attack as mere pretext for U.S. aggression. They respond with a cyber-attack on the electrical grid in the Washington DC area, causing rolling black outs amid a summer heat wave. Several people die.

As the dust-up becomes public, U.S. officials feel pressure to retaliate further. The president orders a missile strike against a Chinese military computer facility. Two-dozen Chinese coders are killed.

What started as a limited cyber-intrusion now threatens to turn into a full-fledged military conflict.

Each element of this hypothetical has occurred in the recent past in some form or another. And it shows how complicated the strategic landscape has become in the cyber age. Unfortunately, our strategic thinking has not kept pace. The United States needs a new strategy for deterrence in cyberspace.

Attribution problems are the chief strategic complication. As former U.S. Deputy Secretary of Defense William Lynn put it: "Whereas a missile comes with a return address, a computer virus generally does not." In congressional testimony, former NSA Director Michael Hayden warned that "[c]asually applying well-known concepts from physical space like deterrence, where attribution is assumed, to cyberspace, where attribution is frequently the problem, is a recipe for failure."

During the Cold War, game theory played a key role in helping policymakers formulate strategies appropriate to the challenges of the nuclear age. Following this tradition, we have taken up General

Hayden's challenge by developing a new game theoretic analysis of deterrence appropriate for the cyber age, one that takes the problem of attribution seriously.

Our analysis yields four key insights:

- Deterrence in cyber-space is fundamentally global and interconnected, not bi-lateral.
- Optimal cyber-deterrence blends aggressive retaliation when attacks are clearly attributable with forbearance when they aren't, rather than across-the-board aggressiveness.
- Retaliatory efforts should be focused on our most deterrable, rather than most aggressive, adversaries.
- Technological improvements in attribution will not always improve deterrence.

A Global Landscape

Our call for a new cyber-strategy comes in the context of other schemes we view as dangerously simplistic. The latest Department of Defense Cyber Strategy focuses on our most belligerent and capable cyber adversaries: China and Russia. The 2018 National Defense Strategy acknowledges North Korea and Iran as rogue nations to contend with and notes the emergence of threats from non-state actors. Nonetheless the 2018 DoD Cyber Strategy emphasizes America's two biggest rivals.

"Our focus will be on the States that can pose strategic threats to U.S. prosperity and security, particularly China and Russia," it reads.

But we cannot be effective in the cyber domain, which contains a much larger number of cyber-capable adversaries, if we narrow our lens in this way.

Consider a country weighing how to respond to a cyber-attack. That country will only retaliate if it is sufficiently confident in its assessment of which adversary is responsible. Of course, it is the anticipation of such retaliation that creates deterrence. This means that an adversary will be more aggressive in cyberspace when it believes the defending country is less likely to reach the confidence threshold necessary for retaliation.

One important input to blame assessment concerns features of the attack itself—e.g., the location of servers, the language and style of malicious code, or the identity of likely beneficiaries. Another input concerns the more general strategic environment. Adversaries believed to be particularly active or capable in the cyber domain will be more suspect following any hard-to-attribute attack. It is this latter fact that makes cyber-deterrence fundamentally global, rather than bilateral.

Suppose some adversary, say China, is believed to have become more aggressive in the cyber domain. Then China is now more suspect whenever a hard-to-attribute cyberattack occurs. China is thus more likely to face retaliation. But if China becomes more suspect, other adversaries, say North Korea, Russia or Iran, must become less suspect. And so these other adversaries are less likely to face retaliation. This reduced risk of retaliation tempts these adversaries to become more aggressive. And so, in cyber-space, if we become worse at deterring any one adversary, we become worse at deterring them all.

This interconnectedness is reflected in several recent cyber-attacks. According to American authorities, a Russian cyber-attack disrupted the opening ceremony of the PyeongChang Winter Olympics. The GRU routed the intrusion through North Korean IP addresses to deflect blame. The North Koreans were an attractive target for this "false flag" operation precisely because they were already highly suspect due the Sony Pictures hack and other cyber operations.

Or consider the "GhostNet" plot, a worldwide infiltration of government and commercial networks, originating in China. A report by the Information Warfare Monitor identifies the Chinese government and military as leading suspects. But it notes that another plausible explanation is "a state other than China, but [operating] physically within China…perhaps in an effort to deliberately mislead observers."

Here we see why attribution problems mean we must think and act globally. If we narrow our focus to China and Russia, we encourage belligerence by other actors. And this increased aggressiveness will create new opportunities in cyberspace for the Chinese and Russians.

Toward a More Effective Cyber Deterrence Strategy

In any deterrence setting, the optimal response after an attack may not be the same as the threat that optimizes deterrence prior to an attack. This familiar problem of credible commitment necessitates that governments articulate and commit to a deterrence doctrine.

Serious discussions are underway about how countries might pre-commit in cyberspace. But, for such pre-commitment to be of use, we must know what the optimal doctrine is. Recent arguments call for a more aggressive retaliatory regime—for instance by declaring governments responsible for cyber-attacks originating in their territory, regardless of the perpetrator. Such calls are consistent with the general theory of deterrence: heightened retaliatory aggressiveness deters more attacks.

But our analysis shows that matters are less clear-cut in the cyber domain. There is a vanishingly small chance that we will engage in, say, nuclear retaliation against the wrong adversary. As William Lynn reminds us, missiles come with a return address. But cyber-attacks do not.

As such, cyber-deterrence doctrine must balance a fundamental trade-off. Committing to a more aggressive retaliatory policy deters more attacks. But it also entails greater risk of mistaken retaliation. Thus, in cyberwarfare, where attribution problems loom large, full deterrence is infeasible. And increased aggressiveness across-the-board is unlikely to be optimal.

The optimal deterrence doctrine for cyber-warfare is more nuanced. We should commit ourselves—through policy declarations, treaties, and standing military orders—to retaliating more aggressively than we otherwise would following clearly attributable attacks. But we should also commit ourselves to retaliating less aggressively following attacks whose attribution is particularly ambiguous. Such forbearance will reduce the risk of erroneous retaliation and dangerous escalatory spirals, with only limited costs for deterrence and security.

And let's return to the global nature of cyber conflict. Despite the 2018 Department of Defense Cyber Strategy's focus on Russia and China, the optimal cyber doctrine doesn't call for increased aggressiveness against our most aggressive adversaries. Rather, it calls for increased aggressiveness against our most *deterrable* adversaries. An adversary is deterrable if its attacks are particularly easy to attribute (e.g., it is technologically limited, other countries aren't trying to mimic it) and if it is particularly responsive to retaliation (e.g., because of its own cyber vulnerability or because of domestic political considerations).

When we improve deterrence against these adversaries, we improve deterrence against all our adversaries, who will have fewer other cyber aggressors to hide behind.

Improving Attribution Doesn't Always Improve Deterrence

If we could attribute cyber-attacks perfectly then deterrence in cyberspace would be no more strategically difficult than in other domains. And indeed, improving our ability to correctly assign responsibility for digital intrusions is a U.S. priority.

The Department of Defense's 2015 official Cyber Strategy has this to say:

> Attribution is a fundamental part of an effective cyber deterrence strategy …DoD and the intelligence community have invested significantly in all source collection, analysis, and dissemination capabilities, all of which reduce the anonymity of state and non-state actor activity in cyberspace….

But matters are more complicated than this statement suggests. The attribution problem actually entails three distinct kinds of potential errors, associated with different challenges.

There is a *false alarm* if a state perceives an attack when no attack occurred. In 2008, a worm gained access to U.S. war planning materials. The prime suspect was Russian foreign intelligence. But others, noting the worm's relative unsophistication, argue it could have ended up on Department of Defense networks without malicious intent. This may, then, have been a false alarm.

There is *detection failure* if a state fails to perceive an attack that did occur. The Stuxnet worm caused centrifuges to malfunction at the Iranian nuclear facility at Natanz for more than a year. The Iranians, though, believed the failures were the result of engineering incompetence or domestic sabotage. This was a case of detection failure by Iran.

And there is *misidentification* if a state assigns responsibility for an attack to the wrong adversary. The hack of Democratic National Committee servers during the 2016 U.S. presidential election was initially attributed to a lone Romanian hacker who went by the moniker Guccifer 2.0. Later, U.S. authorities determined the hack was the work of Russian security agencies who tried to cover their tracks by pretending to be Guccifer 2.0.

Policy arguments regarding the benefits of improved attribution typically do not distinguish between these three dimensions. But our analysis shows that innovations in technology or intelligence that affect different dimensions of attribution can have critically different impacts on deterrence.

Reducing detection failure, for example, has competing effects. On the one hand, improved detection increases our ability to retaliate. On the other hand, improved detection may entail discovering more attacks that are hard to attribute to a specific adversary, increasing concerns about misidentification.

As a consequence of that second effect, such technological progress could backfire—making us more reluctant to retaliate and our adversaries more aggressive. The same sort of logic applies to many kinds of improvements in attribution. For a technological innovation to strengthen deterrence, it must make us more willing to retaliate—for instance, by simultaneously improving detection and identification or by reducing false alarms.

Perhaps most surprisingly, sometimes getting worse at attribution can actually improve deterrence. For instance, if we are reluctant to retaliate following certain types of attack because they are so difficult to attribute to a specific adversary, it is better not to detect them at all. By not detecting such attacks, we make attribution more certain and retaliation more attractive, following those attacks that we do detect. This strengthens deterrence even while worsening attribution.

The hypothetical scenario that began this essay, about stolen military secrets, illustrates the risks of an overly narrow, muscular approach to cyber deterrence. Focusing just on China and Russia, and swinging a big cudgel in response to every cyberattack, tempts other adversaries to be more aggressive in cyberspace, risks retaliation against the wrong party, and can escalate into a potentially catastrophic conflict. A controlled, confident approach that defends the national interest aggressively when the right information is available, while acknowledging that we cannot deter or retaliate against every cyber-attack, is the right path forward in our transformed strategic landscape.

Sandeep Baliga is the John L. and Helen Kellogg Professor of Managerial Economics and Decision Sciences at the Kellogg School of Management, Northwestern University.

Ethan Bueno de Mesquita is the Sydney Stein Professor and Deputy Dean at the Harris School of Public Policy at the University of Chicago.

Alexander Wolitzky is Associate Professor of Economics at the Massachusetts Institute of Technology.

This essay is based on results from a more technical paper on the topic by the authors.