SWP

SWP Comment

NO. 34 AUGUST 2019

Cyber Deterrence is Overrated

Analysis of the Deterrent Potential of the New US Cyber Doctrine and Lessons for Germany's "Active Cyber Defence"

Matthias Schulze

Proponents of active, offensive cyber operations argue that they could have a deterrent effect on potential cyber attackers. The latter would think twice about attacking if a digital counter-attack might be the consequence. The idea that offensive cyber capabilities should have a deterrent effect was one reason why the new US cyber doctrine was adopted in 2018. The same assumption is implicit in the debate about cyber counterattacks ("hack backs") in Germany. Yet these assessments are based on a superficial understanding of deterrence. Cyber deterrence by the threat of retaliation works differently than that of nuclear deterrence. Problems of attribution, displays of power, controllability and the credibility of digital capabilities increase the risk of deterrence failure. Thus, the German cyber security policy would be well advised to increase its "deterrence by denial", cyber security and the resilience of its systems.

Currently, German cyber operators have no legal mandate to conduct disruptive cyber operations outside of German networks in peace time. For this reason, Germany has been debating active cyber defence or "hack backs" for the last few years. Active defence is designed to counter cyber intrusions by striking back at the originator with digital means. These retaliations could be conducted by state entities, not private entities — in stark contrast to the US debate. Proponents of active defence argue that German state hackers should be able to penetrate networks of opponents to stop ongoing cyber attacks in real time, delete data or deactivate computers.

There is another, more implicit argument in the debate for active cyber defence: a cyber attacker could, at least in theory, be deterred from an attack against Germany, if digital retaliation via "hack back" would be the consequence of such behaviour. This mirrors the argument that the presence of offensive cyber capabilities might create a deterrent effect. A similar justification was used for the establishment of the Bundeswehr Cyber and Information Domain service, a functional service of the armed forces in 2016. However, the effectiveness of cyber capabilities for deterrence is the subject of much debate in academic literature. The question thus arises as to whether cyber deterrence by hacking back or by punishment is an appropriate strategy for Germany.

Stiftung Wissenschaft und Politik
German Institute for International and Security Affairs

Deterrence

Deterrence is the potential use or threat of punishment to achieve a change in behaviour of an opponent. Deterrence is based on the formula that the offensive behaviour (X) of an attacker (A) can be changed by the defender (D), if he is threatening him with negative consequences (Y). The logical formula for deterrence is: Do not do X, because otherwise consequences Y will follow. The cost of Y must outweigh the gains to be expected from an attack X. This form of deterrence always contains an element of coercion and is therefore called "deterrence by punishment". It differs from "deterrence by denial", which aims to increase the cost of attacks by hardening systems and increasing resilience so that attacks no longer seem worthwhile. Unless otherwise stated, I mean deterrence by punishment when speaking of deterrence in this paper.

In order for deterrence to work, at least three conditions must be met:

- The threat of consequences must be clearly communicated and understood by all parties ("signaling").
- Both actors must have as complete information as possible about the capabilities, intentions and ideally the thought processes of their counterparts in order to be able to rationally assess costs and benefits.
- The threat of punishment must be credible, i.e. technically feasible and backed by political resolve.

Successful deterrence requires the threat of punishment to be communicated in a clear, audible and, above all, credible manner. Deterrence is considered successful if A does *not* perform an action, i.e. a cyber-attack. The causality of a non-event cannot be proven logically. One can never say exactly whether it was the threat of punishment that led to the change in behaviour, or whether there were other reasons for it. Consequently, deterrence is sometimes considered to be a myth in academic literature.

Deterrence is based on Rational Choice Theory. The assumption is that actors weigh the costs and benefits of their actions and rationally choose the less costly option. The theory has been criticized because actors never have all objective information and can therefore never fully assess the consequences of their actions. Furthermore, they often act irrationally, rely on bounded rationality, or act according to norms or habits. Deterrence works only in the head of the attacker, where one has no insight. So it is ultimately a guessing game: "I believe that you believe that I believe" and so on. The logical problem with all deterrence theories is that you never know if deterrence works, until it fails.

Cyber Deterrence

Transferring a deterrence strategy to the cyber domain is regarded as problematic by cyber security researchers. Nuclear deterrence was assumed to be successful due to unique conditions, i.e. the particularities of the bi-polar world and the extraordinary damage potential of nuclear weapons, which made defence strategies less feasible. In the bipolar world of the Cold War, deterrence was symmetrical and applied by roughly equally strong actors who were able to assess their motives sufficiently well. Cyber deterrence is multipolar and takes place between asymmetric opponents. Cyber capabilities are mostly opaque and easily proliferate. In this respect, cyber deterrence can fail more easily and is therefore not a reliable policy option.

Attribution Problem

Successful attribution is the most important prerequisite for deterrence, as it provides legitimacy and the threat of punishment with a certain strategic gravitas. However, it is often unclear who is behind cyber incidents. Consequently, no one can be identified who can be threatened with punishment. The attribution problem describes the difficulty of apportioning responsibility

of cyber attacks to an actor who has not previously communicated his intention and left no confession.

The attribution problem affects both sides: When A cyber attacks D, D does not automatically know that it was A. If D retaliates digitally, again A does not necessarily know that it was D. There is barely a target in digital space that is attacked by only one actor. Misperceptions are therefore quite common. There is also the risk that attackers may act under a false flag or claim to be responsible for attacks they did not carry out. In escalating geopolitical conflict situations, however, the role of the attribution problem is probably overrated. If, for example, servers are flooded in South Korea during a conflict episode with North Korea, it is easier to see who benefits from this ("cui bono") than it is with covert espionage operations. For effective deterrence, however, attribution must be incontestable, accurate and immediate. The more time that elapses between incident and attribution, the less legitimate a cyber retaliation by D.

Demonstration Problem

An attacker must be able to weigh up the costs of a potential punishment by D. Thus, A must be able to assess the damage potential of D's cyber capabilities. For this very reason, military parades display kinetic weapons to the world and weapons tests are conducted for the whole world to see. This transparency principle, however, does not readily apply to cyber capabilities. Demonstrating of cyber capability for reasons of damage threat jeopardizes the functioning of the capability. If a defender knows about the attack vector, he can adapt, which then makes an attack less useful. Offensive cyber abilities follow the law of diminishing returns: any deployment of ability increases the chances that it will be less effective in the future.

A low-threshold Distributed Denial of Service (DDoS) attack may succeed the first time. However, if the attacker knows that retaliation is imminent, he or she can take critical systems off the network as a precaution, or redirect the harmful network traffic. DDoS attacks are therefore only of limited use as a potential punishment. The same problem exists with 0-day capabilities, i.e. attacks that are based on unknown and therefore unpatched vulnerabilities. The more frequently they are used, the greater the probability that they will be exposed and thus made available to the entire world. With a patch for the vulnerability, the capability loses its effectiveness.

This has two implications: 0-day capabilities cannot be credibly demonstrated without compromising their effectiveness. They are therefore only suitable for threatening punishment to a limited extent. The exception would be if an attacker had several hidden backdoors for accessing an enemy system. Then 0-day attacks could be used for "signaling". Second, a defender can repurpose a published 0-day ability and direct it against the attacker. This suggests the risk of blowback for any attacker A, whether by D, or by any third party that repurposes the malware.

Proportionality and Appropriateness

Deterrence fails if the threat of punishment is not considered credible. Deterrence failure often leads to the use of capabilities and thus escalation. This raises questions about the proportionality, effectiveness and accuracy of cyber retaliation capabilities. How much objective damage must be inflicted so that A considers the costs of further offensive action to be too high? How does D know whether A considers threats against certain assets to be particularly painful or not? A and D most likely have different perceptions about what assets are considered especially sensitive. These different perceptions make proportional reactions difficult. There is no international consensus on how proportional cyber retaliation might be conceived. Thus there is an increased risk of escalation.

The damage caused by cyber retaliation must be appropriate. If the damage threat-

ened by D is too great, the probability of a renewed retaliation by A increases. It is well researched in political science, that escalation spirals are often a consequence if a retaliation is perceived as inappropriate or too painful. In these cases, deterrence fails. If the threat of punishment is considered not costly enough and thus not credible, deterrence does not work either. Determining the correct measure is highly complex and also a function of the attribution problem: the lower the chance of being caught, the greater the threat of punishment by D must be, if A is to be convinced that an attack is not worth the potential cost. Another issue is that particularly costly assets are usually well protected, which makes effective retaliation harder.

Lack of Controllability

The damage potential of cyber capabilities is unreliable and difficult to control. It is complicated, although not impossible, to limit cyber capabilities to one target and to avoid collateral damage, for example in uninvolved third countries. This is particularly true in time-critical situations. The effectiveness and thus the exact damage potential of cyber capabilities are often difficult to determine in advance. The potential damage is largely determined by the configuration of the target system. In this respect, it is often impossible to anticipate how long a cyber attack can disrupt a system, for instance.

This fact complicates the proportional and controlled use of such capabilities. This in turn increases the risk of deterrence failure. Even attacks such as Stuxnet (2010), which were carefully tailored to specific targets, also infected other systems worldwide. Collateral effects such as WannaCry or NotPetya (both 2017) are habitual in cyber conflicts. No one can realistically estimate where else a certain system configuration is in use.

On the other hand, threat of punishment can be made too specific. If, for example, D is about to respond to a cyber attack on a dam by A with a retaliatory strike on a dam owned by A, A can take this off the grid as a precaution. It is difficult to find the right measure for potential damage that is neither too precise nor too vague, especially as the risk of deterrence failure is high. Furthermore, the risk of escalation increases in asymmetric contexts. This makes cyber capabilities seem unreliable as a deterrent.

High and Low-Level Deterrence

There is no international consensus as to what cyber activities can be considered legitimate for deterrence (political vs. economic espionage vs. sabotage). Depending on the intensity of the activities, the chances of success for deterrence may vary. High level deterrence is aimed at preventing cyber activities that reach the threshold of an armed attack. This includes the worst-case scenario of a digital surprise attack on strategic infrastructures, in which people die and high-grade physical destruction is the result ("digital Pearl Harbor"). Such an event has never happened in the more than thirty-year history of cyber-conflicts. The reason is that its consequences could not be measurable, costs would be too high, and an attacker would probably face blowback effects.

First, such an attack would most likely be considered an act of war under international law and would legitimize, for example, acts of (collective) self-defence. Such a cyber attack would therefore probably escalate into a physical conflict, which is why states refrain from these activities in peacetime. Secondly, due to the interdependent and highly networked Internet infrastructure, it cannot be realistically guaranteed that one's own systems would not be similarly affected. In view of this, states have no interest in carrying out such strategic attacks in peace time, unless they can really gain something politically. Here, an implicit norm of restraint is effective, which is also noticeable in various international norm-setting bodies. In other words, deterrence can also work through norms that put a taboo on inappropriate behaviour.

However, this reluctance does not exist in the case of low-level incidents, below the threshold of an armed attack. States deliberately design their cyber activities in such a way that they remain below this threshold and thus do not have an escalating effect. This category includes cyber espionage, hybrid measures, cybercrime, hacktivism and vandalism, which account for a large proportion of all cyber activity. It is considered unlikely that deterrence will be effective in low-threshold incidents such as espionage. There is a high likelihood of not being caught, especially since states are not interested in punishing espionage, from which they themselves benefit.

Non-State Actors

Low-level cyber activities are also committed by non-state actors. This is a major difference from deterrence in the nuclear age, where only states possessed nuclear capabilities. The spectrum of actors ranges from script kiddies with low level skills to cyber criminals with medium abilities to cyber mercenaries with considerable capabilities. In addition, there are so-called proxy actors who attack targets either independently, or on behalf of a state.

Deterrence only works if the motivation, interests, skills and return address of the opponents are known. With many "advanced persistent threats" much of this information remains opaque. Therefore, they cannot be effectively deterred. Theoretically, an effective deterrence policy would need to be tailored to each opponent among the thousands of cyber actors. This is impossible even for great cyber powers.

It is well known from terrorism research that deterrence by punishment works, if at all, only against states, but not necessarily against non-state actors. Here, deterrence can produce a converse effect: the use of repressive force to combat terrorism often leads to more terrorism due to the perceived injustices. The same can be observed in digital space. Not even offensively dominant states such as the USA are in a position to deter cyber attacks by non-state or state actors such as Russia or China. Deterrence of non-state actors follows the logic of criminological deterrence, which aims to reduce the frequency and intensity of incidents without being able to prevent them altogether.

There is another problem with non-state actors: not all of them act according to the same rational principles to which states, presumably, would act. Hackers, for example, are not necessarily driven by rationale, but also by cognitive and normative motivations, such as the desire to gain fame and have fun ("Lulz").

Credibility and Escalation

The threat of punishment not only needs to induce an accurate estimate of the expected costs, it must also be credible. If A does not believe that D, firstly, is technically capable of causing precisely measured costs with digital means or, secondly, lacks the political will or resolve to endure the risk of escalation, deterrence fails.

The credibility problem is even greater in cyber conflicts. Intentions and political will are often unclear, as much of government cyber activity is carried out by intelligence services and falls under cyber espionage. Thus, intention and political will remain hidden in many cyber-incidents. The intrusion into systems for espionage or sabotage purposes cannot be clearly distinguished from one another by the defender. This increases the risk that D perceives a relatively harmless act of espionage as an attempt at sabotage, and thus overreacts. Furthermore, states are unable to objectively assess their relative cyber-power. Cyber capabilities cannot be counted like tanks or warships. As "Rational Choice Theory" deterrence requires as complete information as possible, which also includes an assessment of relative strength. This fails because of the secrecy and dual-use nature of cyber capabilities, which can be used for offensive and defensive purposes.

Moreover, not all states are politically willing to engage in a "tit for tat" escalation dynamic of mutual retaliatory strikes. In

game theory, such conflicts are referred to as "chicken games". In the classic scenario, two actors race directly towards each other in the car; the one who swerves first is the "chicken", the coward. In democracies, the electorate usually does not support aggressive foreign policy. Therefore, the executive often has less leeway to credibly threaten punishment. However, credibility also depends on past decisions and the reputation of a government. If the government has reacted hesitantly to aggression in the past, their future threats of punishment are less credible.

The problem with gradual escalation in cyberspace is that the damage of the retaliatory attack must be somewhat higher than that of the previous attack. Since it is difficult to determine proportionality, there is a risk of collateral damage. It is unclear how escalation dynamics function in cyberspace. There is no clear consensus among scholars about whether cyber capabilities can reach a similar level of escalation as physical weapons, or whether they are in principle de-escalatory. Some commentators argue that digital means tend to limit escalation because physical effects are difficult to produce, and the damage potential is more limited. Empirically, escalation is the most likely outcome of a deterrence policy that predominantly relies on the use of offensive means.

Deterrence and "Persistent Engagement" in the US Doctrine

Deterrence is thus not easily transferable to the digital domain. Hawks and national security advocates, however, disagree and believe that, in case of doubt, the possession of fearful cyber capabilities produces deterrent effects. They advocate a stronger offensive, because although the US is a formidable cyber power, it could not deter Russia from influencing the 2016 US presidential election with cyber capabilities. In response to this deterrence failure, the Pentagon introduced a new cyber doctrine in 2018. This contains new concepts such as "defending forward", "persistent engagement" and "preparation of the battlefield". The doctrine gives the US Cybercommand greater scope for offensive action, for which no presidential authorization is required.

Defending forward means that networks are no longer defended in one's own perimeter or territory, but on the systems of potential attackers. This potentially includes unwitting third parties worldwide. Attacks against opponent systems are primarily used to gain intelligence in order to detect enemy attacks and burn capabilities at an early stage.

"Persistent engagement" means binding enemy forces by permanently exposing them to attacks by American hackers. Opponents would constantly have to defend themselves against American intrusion attempts so that — according to the theory — they no longer have resources for their own offensives. Since no other state has such large personnel resources as the USA, the costs for attackers would be increased in this way. The doctrine clearly mentions China and Russia as potential targets for these measures.

Defending forward and persistent engagement are operational strategies that by themselves are not designed for strategic deterrence. However, it can be argued that the third concept, the "preparation of the battlefield", might have deterrent effects.

Opponent networks are to be penetrated in order to implant so-called back doors or logic bombs which can be exploited in future conflicts. A logic bomb is malware that lurks undetected in a network until it is activated at a later point in time. This implies a concrete threat of punishment. An opponent would then always have to ask themselves whether they have uncovered all the attack vectors of the Americans or whether they overlooked a hidden back door in their own network. In view of this uncertainty, attackers could refrain from serious cyber attacks, for example against critical infrastructures. Russia has recently complained vociferously about attempts by American hackers to penetrate the Russian power grid in order to implant

backdoors. The Kremlin also warned against an escalation in the cyber area. This is an indication that the USA's new cyber doctrine, which is even more offensive than its predecessor called "active defence", might be fuelling escalations. Whether it does so empirically remains to be seen.

"Persistent engagement" was applied during the Midterm Elections of 2018. A central hub of low-level Russian cyber activity, the "troll factory" or Internet Research Agency in St. Petersburg, was temporarily disrupted. However, it resumed its activities shortly afterwards. Tactically, the operation may have been a success. However, it is doubtful whether this form of deterrence has a strategic, i.e. long-term effect. It is to be expected that other cyber powers will now also invest more offensively and train more personnel in order to withstand or outmanoeuvre such "persistent engagement".

The result would be an intensified arms race with the aim of always being able to mobilize more cyber forces than its rival. It remains to be seen whether persistent engagement will work against more than a handful of opponents at the same time. Low-threshold attackers cannot be stopped in this way either.

Persistent engagement is a NOBUS strategy — nobody but US — and thus cannot be easily replicated by other cyber powers. However, if all cyber powers were to pursue such a doctrine and start placing back doors everywhere, global cyberspace would be highly volatile. Backdoors are not exclusive and can potentially be exploited by any knowledgeable attacker. The cost of such an offensive policy for collective security would probably be higher than the theoretical gain in national security. The new doctrine thus goes far beyond the concept of "active cyber defence" of the Obama era. The concept was to react offensively to cyber attacks, but only to stop them at their source. This is also the concept that the German government is currently considering in a modified form.

Cyber Deterrence by German Active Defence?

Whether the mere possession of German cyber offensive capabilities would have a deterrent effect is doubtful. All the problems of attribution, demonstration, proportionality and controllability of cyber retaliation described above still apply. Furthermore, it is hard to believe that Germany would be prepared to enter into a dynamic of escalation in cyberspace and then possess the necessary resolve. The culture of restraint in foreign and security policy is still very pronounced. The population is critical of a more active foreign policy or the assumption of greater responsibility. This is particularly true if the use of force is involved, whether physical or digital.

Germany would probably have a credibility problem, if it were to adopt a deterrence-by-punishment posture. A strong opponent would want to test whether Germany is politically prepared to use active cyber defences as a deterrent and is willing to endure the consequences of an escalation. So far, Germany lacks a political strategy on how to deal with such a situation. It would have to be tailored to all relevant cyber threats and include the aforementioned elements of threat communication as well as measures to provide proportional and effective cyber reaction tools. Additionally, political will is required to use cyber capabilities as a form of punishment, even in the face of a probable escalation dynamic. Whether this actually exists is doubtful.

Since an escalation strategy and political resolve for deterrence by punishment does not exist, "deterrence by denial" is a better strategy for Germany. This conclusion can be derived from the deficits of deterrence by retaliation itself. It fails inter alia because targets are too easily attackable. The bottom line is that it is always cheaper for the attacker to exploit weaknesses than not to do so.

The first step towards an effective deterrent system should therefore be to increase cyber security and resilience in order to

make cyber attacks more costly. Of course, deterrence by denial faces several problems itself, so this will not be easy. As a second step, deterrence that accompanies foreign policy measures should be extended. There is much to suggest that deterrence, if at all, only works in concert with other measures — at best within the framework of an international cyber regime that does not yet exist. This would include international diplomacy, deterrence through norms or international interdependence or entanglement, but also through regimes and organisations that subject state behaviour to rules. The efforts of cyber foreign policy should be intensified in this direction. However, this is a long way off.

Cyber-conflicts are largely unregulated. Established norms for appropriate behaviour and red lines do not yet exist. Consequently there is a high risk that deterrence will fail and trigger an escalation dynamic. Germany should therefore consider whether it wants to participate in this game, and whether it is prepared to endure any negative consequences. Cyber security by resilience is in any case the more long-lasting strategy, since it works against all opponents in the same way, and does not need to be tailored to specific opponents.

Summary

The existence of offensive cyber capabilities alone does not act as a deterrent, especially if it is not credibly communicated that there is a willingness to use them. There are many pitfalls that make deterrence by punishment an ineffective policy concept with many risks. The risks of deterrence failure are more prevalent than in the analogue world. Deterrence by punishment is most likely a strategy doomed to fail.

If even the more active cyber powers like the US regularly fail with cyber deterrence, then a German cyber deterrence policy — due to the traditional restraint in foreign and security policy — cannot be expected to be successful either. As long as Germany has no escalation strategy and is not prepared to endure the possible consequences of an offensive cyber deterrence policy, this approach should be avoided. Instead, German policy should continue to focus on "deterrence by denial" by hardening systems and building resilience.

Dr Matthias Schulze is Associate in the International Security Division at SWP.

SWP
Stiftung Wissenschaft und Politik
German Institute for International and Security Affairs

Ludwigkirchplatz 3—4
10719 Berlin
Telephone +49 30 880 07-0
Fax +49 30 880 07-100
www.swp-berlin.org
swp@swp-berlin.org

ISSN 1861-1761
doi: 10.18449/2019C34

(English version of SWP-Aktuell 39/2019)