November 2021 · Dr. Sven Herpig

Active Cyber Defense Operations

Assessment and Safeguards

An analysis supported by the Transatlantic Cyber Forum

Stiftung Neue Verantwortung

Think Tank at the Intersection of Technology and Society

Executive summary

The policy world has debated active cyber defense for many years. However, many of those discussions have not been concluded, at either the international or national level. Active cyber defense—the implementation of technical measures to mitigate, neutralize and attribute a malicious cyber operation or campaign[1]—warrants this thorough analysis due to its inherent risks.

Whether the implementation of a measure is covered by an existing legal framework is certainly a good starting point. However, existing legal frameworks are applied to new areas, such as active cyber defense, for which the frameworks were not designed. Especially in the absence of specialized legal frameworks, not everything that may not be illegal should be pursued. Thus, how do we assess whether active cyber defense measures should be implemented?

Definitions of active cyber defense vary broadly, as do the technical measures that fall under those definitions. Creating an exhaustive list of measures and deciding which ones are useful and which are not seems like a futile task. That is why this paper suggests an overall framework for assessing whether measures should be implemented. The following criteria are included in this framework:

- Goals and success (purpose);
- Type, space and target of effect (effect);
- Government lead agency and cooperativeness of the stakeholders (actors);
- Attribution and time (timing);
- Escalation, automation, frequency, costs and collateral consequences (operations).

In addition to the assessment of individual measures, risks should further be mitigated by establishing safeguards applicable to every active cyber defense measure. Safeguards that should be implemented are as follows:

- Define and limit the scope;
- Establish a national legal framework that includes transparency, oversight and impact assessment;
- Set up guidelines for tools and services;
- Apply international law;
- Consider public interest;
- Adapt confidence-building measures.

1 For the full definition, see section 2.

It is important to move the policy discourse on active cyber defense forward. However, national and international cybersecurity will always depend on improving IT security and resilience. Active cyber defense can occasionally supplement IT security and resilience but will never substitute them. The assessment framework and safeguards ensure that in cases in which active cyber defense operations are required to fill a gap in protecting the state from malicious cyber operations and campaigns, they do so without causing more harm than good.

Table of Contents

Executive summary 2

1. Introduction 6

2. Definition 11

3. Criteria 15
3.1. Purpose 18
3.1.1. Goal 18
3.1.2. Success 18
3.2. Effect 19
3.2.1. Type 19
3.2.2. Space 19
3.2.3. Target 21
3.3. Actors 22
3.3.1. Government lead agency 22
3.3.2. Cooperativeness 23
3.4. Timing 23
3.4.1. Attribution 23
3.4.2. Time 24
3.5. Operations 25
3.5.1. De- and escalation 25
3.5.2. Automation 26
3.5.3. Frequency 27
3.5.4. Costs 27
3.5.5. Collateral consequences 28

4. Safeguards 29
4.1. Define and limit the scope 29
4.2. Establish a national legal framework 30
4.2.1. Require impact assessments 31
4.2.2. Implement oversight 31
4.2.3. Create transparency and auditability 32
4.3. Set up guidelines for tools and services 33
4.4. Apply international law 34
4.5. Consider public interest 35
4.6. Adapt confidence-building measures 35

5. Applications 37
5.1. Removal of the Hafnium web shells 37
5.1.1. Background 37
5.1.2. Criteria 38

5.1.3. Safeguards 41
5.1.4. Assessment 42
5.2. Disabling of Emotet malware 43
5.2.1. Background 43
5.2.2. Criteria 44
5.2.3. Safeguards 47
5.2.4. Assessment 49

6. Recommendations 51

Acknowledgment 53

1. Introduction

The idea of states and even private entities[2] implementing active cyber defense measures to neutralize, technically attribute or mitigate the impact of ongoing offensive cyber operations and campaigns, including everything from crime to espionage and subversion, has been debated for many years without clear consensus. This paper focuses exclusively on government-led activities in this area and proposes a framework against which past and future active cyber defense operations can be assessed.

Most often referred to as *active cyber defense*, the discussions have led to an intellectual turf war between three, with the latter two sometimes overlapping, "schools of thought" on active cyber defense: 1) those who claim that relying on passive defensive methods is not enough and that offensive methods are needed based on structural features of the strategic environment[3]; 2) those who argue that most of the basic defensive measures have not really been implemented across sectors and if they had been, they would have sufficed[4]; and 3) those who see active cyber defense as yet another euphemism used by governments to legitimize offensive cyber operations and refer to them—somewhat mockingly—as *hacking back*[5]. In most cases, this debate falls into a false binary between allowing highly intrusive cyber operations or maintaining the restrictions of the status quo. The advocates of more intrusive defensive operations and those categorically against active cyber defense often fail to acknowledge that this is a spectrum of diverse measures that should each be examined on its own merits. Moreover, these debates do, for the most part, not even cover the international (legal) dimension,[6] potential spillover effects and challenges of technical attribution. Healey, Jenkins and Work concluded that

> *"[t]he disconnects between policy communities and the operators and researchers engaged in the day-to-day fight on the wire have meant that in many cases, well-intentioned thinkers on both sides have been effectively talking past each other when discussing concepts of operation, desired end states and perceived drawbacks".*[7]

2 Brad D. Williams (2021), Proposed 'Hack-Back' Bill Tells DHS To Study Allowing Companies To Retaliate, Breaking Defense and Tom Graves (2019), H.R.3270 - Active Cyber Defense Certainty Act, U.S. Congress and Center for Cyber and Homeland Security (2016), Into the Gray Zone - The Private Sector and Active Defense Against Cyber Threats, The George Washington University and Patrick Lin (2016), Ethics of Hacking Back, California Polytechnic State University

3 For example, Ciaran Martin (2021), Offensive cyber in the age of ransomware, Offensive Cyber Working Group

4 For example, Jack Goldsmith and Robert D. Williams (2018), The Failure of the United States' Chinese-Hacking Indictment Strategy, LAWFARE

5 For example, Manuel Atug (2021), Hackback statt Cyberresilienz, Heise Magazine

6 For example, Henning Lahmann (2020), Unilateral Remedies to Cyber Operations, lehmann media and Ashley Deeks (2020), Defend Forward and Cyber Countermeasures, Hoover Institution and Janine Schmoldt (2020), Hacking Back aus völkerrechtlicher Perspektive, Defensive Con 2020

7 Jason Healey, Neil Jenkins and JD Work (2020), Defenders Disrupting Adversaries: Framework, Dataset, and Case Studies of Disruptive Counter-Cyber Operations, CyCon 2020

At the highest political level, however, events often drive policy-making. In July 2021, U.S. President Joe Biden signaled a willingness to "attack the actual servers that are used [...] to carry out these ransom attacks in Russia."[8]

The varying perspectives on active cyber defense are compounded by the absence of consensus on what constitutes an activity that is considered active cyber defense. Although this complicates the debate further, it is not uncommon in the field of cybersecurity policy or any other policy field that definitions vary widely, especially across countries. Based on the work of several experts, this paper develops and uses the following definition of active cyber defense and applies it throughout the analysis for consistency and clarity.

Active Cyber Defense: one or more technical measures implemented by an individual state or collectively, carried out or mandated by a government entity with the goal to neutralize and/or mitigate the impact of and/or attribute technically a specific ongoing malicious cyber operation or campaign.

When enumerating measures that are considered active cyber defense, one can come up with a broad range that includes everything from canary tokens and traffic redirection to *nematodes*[9] and benevolent wiper software.10 Advancing a response to this challenge likely results in a nuanced assessment rather than a one-size-fits-all solution.

As is often true, the reality for active cyber defense is more evolved than theoretical discussions. Unfortunately, reality is sometimes even more evolved than legal and policy frameworks. In 2021, at least two highly noteworthy active cyber defense operations were conducted.[11] One was the removal of web shells on IT systems that were left behind through the exploitation of ProxyLogon by the Hafnium group. This effort was led by the U.S. Federal Bureau of Investigation (FBI).

8 The White House (2021), Remarks by President Biden Before Air Force One Departure, The White House
9 SecurityFocus (2006), Good worms back on the agenda, Security Focus and
Vesselin Bontchev (1994), Are "Good" Computer Viruses Still a Bad Idea?, personal website
10 Sven Herpig (2018), Aktive Cyber-Abwehr / Hackback, Stiftung Neue Verantwortung
11 For a dataset of active cyber defense operations that partially fall under the definition used in this paper and are referred to as *disruptive counter-cyber operations* by the authors, see Jason Healey, Neil Jenkins and JD Work (2020), Defenders Disrupting Adversaries: Framework, Dataset, and Case Studies of Disruptive Counter-Cyber Operations, CyCon 2020

FBI Active Cyber Defense Operation (Hafnium)[12]**:** On March 2, 2021, Microsoft disclosed that a "state-sponsored threat actor" (Hafnium) operating from China had "engaged in a number of attacks using previously unknown exploits targeting on-premises Exchange Server software."[13] This and other malicious campaigns were able to intrude into and install web shells on the servers via ProxyLogon vulnerabilities.[14] At the same time, Microsoft released updates permitting the patching of these exploits. The update was followed up with a Joint Advisory issued by the FBI and the Cybersecurity and Infrastructure Security Agency (CISA) on March 10[15]. Despite the availability of the patches and the advisory, "hundreds of vulnerable computers in the United States" were not patched, and the respective companies did not remove the web shells until the end of March.[16] The FBI requested a search-and-seizure warrant that would enable the agency to remotely remove the web shells, because the agency believed "that the owners of the still-compromised web servers did not have the technical ability to remove them on their own and that the shells posed a significant risk to the victim"[17] and more generally, "threaten[ed] the national security and public safety of the American people and our international partners."[18]

The other operation was the internationally coordinated takedown of the Emotet botnet, in which Germany's Federal Criminal Police Office (BKA) played a leading role in the malware-disabling part.

12 For more information about the removal of Hafnium web shells and corresponding sources, see subsection 5.1.
13 Tom Burt (2021), New nation-state cyberattacks, Microsoft
14 Tara Seals (2021), Microsoft Exchange Servers Face APT Attack Tsunami, Threatpost and Catalin Cimpanu (2021), FBI operation removed web shells from hacked Exchange servers across the US, The Record
15 Cybersecurity & Infrastructure Security Agency (2021), FBI-CISA Joint Advisory on Compromise of Microsoft Exchange Server, Cybersecurity & Infrastructure Security Agency
16 U.S. Department of Justice (2021), Justice Department Announces Court-Authorized Effort to Disrupt Exploitation of Microsoft Exchange Server Vulnerabilities, U.S. Department of Justice
17 Lawrence Abrams (2021), FBI nuked web shells from hacked Exchange Servers without telling owners, Bleeping Computer
18 U.S. Department of Justice (2021), Justice Department Announces Court-Authorized Effort to Disrupt Exploitation of Microsoft Exchange Server Vulnerabilities, U.S. Department of Justice

BKA Active Cyber Defense Operation (Emotet)[19]: On January 27, 2021, Europol announced that the "world's most dangerous malware Emotet" had been disrupted. The announcement was preceded by a collaborative multinational effort, also known as Operation Ladybird, that included the participation of authorities in the Netherlands, Germany, the United States, the United Kingdom, France, Lithuania, Canada and Ukraine under the coordination of the European Union Agency for Law Enforcement (Europol) and the European Union Agency for Criminal Justice Cooperation (Eurojust).[20] Operation Ladybird allowed government agencies to "gain control of the infrastructure and [take] it down from the inside"[21] and place the malicious software in quarantine within infected machines[22]. Then, the authorities also adjusted "communication parameters of the software [...] in a way that the victim systems no longer communicate with the infrastructure of the offenders but with an infrastructure created for the seizure of evidence."[23]

Based on publicly available information about the achievements and collateral consequences, both operations can be considered successful. Although the FBI measures were criticized as a "drastic step"[24] and "should be considered harmful,"[25] they appear to have a firm legal grounding.[26] In contrast, some policy and legal scholars argue that the BKA's measures were possibly illegal, as the implementing federal agency is not permitted to conduct active defense operations against cybercrimes.[27] These assessments also reflect the policy discussions in the two countries. The United States amended Criminal Procedure Rule 41 in 2016 to enable these kinds of measures for law enforcement[28] and adopted a somewhat preemptive stance to *defend forward* in 2018 within the military domain.[29] However, the German policy

19 For more information about the disabling of Emotet malware and corresponding sources, see subsection 5.2.
20 European Union Agency for Law Enforcement Cooperation (2021), World's Most Dangerous Malware Emotet Disrupted Through Global Action, Europol
21 Andy Greenberg (2021), Cops Disrupt Emotet, the Internet's 'Most Dangerous Malware', WIRED
22 Sergiu Gatlan (2021), Emotet malware nukes itself today from all infected computers worldwide, Bleeping Computer
23 Sergiu Gatlan (2021), Emotet malware nukes itself today from all infected computers worldwide, Bleeping Computer
24 Brian Barrett (2021), The FBI Takes a Drastic Step to Fight China's Hacking Spree, WIRED
25 Ed Amoroso and Randal S. Milch (2021), Hack-to-Patch by Law Enforcement Is a Dangerous Practice, JUST SECURITY
26 Alex Iftimie (2021), No Server Left Behind: The Justice Department's Novel Law Enforcement Operation to Protect Victims, LAWFARE
27 Sven Herpig and Dennis-Kenji Kipker (2021), German Emotet takedown in the legal gray zone, Stiftung Neue Verantwortung
28 Eversheds Sutherland (2016), Amendment to Criminal Procedure Rule 41 Impacts Data Privacy in U.S. and Abroad, JDSUPRA
29 U.S. Department of Defense (2018), Summary - Department of Defense - Cyber Strategy 2018, U.S. Department of Defense and Erica D. Borghard (2020), Operationalizing Defend Forward: How the Concept Works to Change Adversary Behavior, LAWFARE

debate[30] ceased abruptly in 2020 after intense expert-level discussions due to disagreement among the ruling parties.[31] Although it is not uncommon for American and German perspectives regarding law enforcement, intelligence or military deployment of cyber means to vary widely, Germany is not alone in treading carefully regarding active cyber defense. Japan, for example, has also raised numerous concerns in that policy field.[32] In contrast, Israel, for example, "is not shy when it comes to hacking back."[33] Therefore, it is entirely possible that the countries' individual active cyber defense policy debates reflect—at least to a degree—their general strategic culture.

Although the perspectives on active cyber defense vary widely across governments, the objective is the same: protecting the interests of the state, its industry and its citizens in and through cyberspace. Therefore, a nuanced approach to active cyber defense may potentially narrow the current gap between the three schools of thought by offering a viable middle ground. In that spirit, April Falcon Doss argued that

> *"[o]pinion polls and position papers won't affect the underlying legality of the warrant or operation, but they may serve as significant indicators of the extent to which similar operations are perceived as legitimate and appropriate cyber defense tools in the future."*[34]

Outline: This paper focuses on government-led activities. It proposes a definition for active cyber defense operations. Instead of listing tools and measures, the paper lists features and characteristics. Therefore, the analysis discusses the criteria for assessing the effectiveness and risks of an operation. The paper suggests necessary safeguards to mitigate associated risks. Together, the criteria and safeguards constitute a framework for assessing active cyber defense operations, which are tested with two case studies: removal of the Hafnium web shell and disabling of Emotet malware.

30 Sven Herpig et al. (2020), Aktive Cyberabwehr/ Hackback in Deutschland, Stiftung Neue Verantwortung
31 Martin Knobbe and Wolf Wiedmann-Schmidt (2020), Wie die SPD ein Gesetz zum Cyber-Gegenschlag verhinderte, SPIEGEL
32 Nori Katagiri (2021), From cyber denial to cyber punishment: What keeps Japanese warriors from active defense operations?, Asian Security
33 Sven Herpig, Robert Morgus and Amit Sheniak (2020), Active Cyber Defense- A comparative study on US, Israeli and German approaches, Konrad-Adenauer-Stiftung
34 April Falcon Doss (2021), We're From the Government, We're Here to Help: The FBI and the Microsoft Exchange Hack, JUST SECURITY

2. Definition

When a concept that exists in the kinetic world evolves into its counterpart in the cyber domain, it sometimes helps to look at the pre-cyber meaning of the term as a starting point to coin a term. However, that might not suffice for active cyber defense. George Washington University's Center for Cyber and Homeland Security states that

> *"Active Defense is a term that has been in use within the national security and defense communities for a number of decades. Since its origins in the Department of Defense and its later application to the cyber domain, it has taken on a whole host of meanings. Today, the legacy of its various and evolving interpretations obscures the utility of a term in a sea of conflicting definitions."*[35]

Although, unsurprisingly, there is no generally agreed-upon definition for active (cyber) defense, conceptual discussions of techniques involved in active cyber defense and related terms such as *defend forward*, *disruptive counter-cyber operations* or *proactive* (active) cyber defense have significant overlap when it comes to key features.[36]

This paper uses the following working definition for an active cyber defense operation: *one or more technical measures implemented by an individual state or collectively, carried out or mandated by a government entity with the goal to technically neutralize and/or mitigate the impact of and/or attribute technically a specific ongoing malicious cyber operation or campaign.*

First, active cyber defense is, as the name suggests, considered to be defensive and not offensive at the strategic level.[37] Nevertheless, tools and measures used at the tactical and operational levels can be intrusive and also be used for offensive purposes.[38] Therefore, key enablers of active cyber defense operations may well be offensive tools and measures. However, although this abstract differentiation may make sense at the conceptual level, in practice the lines are blurry.

35 Center for Cyber and Homeland Security (2016), Into the Gray Zone - The Private Sector and Active Defense Against Cyber Threats, The George Washington University

36 For example, Center for Cyber and Homeland Security (2016), Into the Gray Zone - The Private Sector and Active Defense Against Cyber Threats, The George Washington University and Jack Goldsmith and Alex Loomis (2021), "Defend Forward" and Sovereignty, Hoover Institution and Thomas Reinhold und Matthias Schulze (2017), Digitale Gegenangriffe, Stiftung Wissenschaft und Politik and Jason Healey, Neil Jenkins and JD Work (2020), Defenders Disrupting Adversaries: Framework, Dataset, and Case Studies of Disruptive Counter-Cyber Operations, CyCon 2020 and Sven Herpig (2018), Aktive Cyber-Abwehr / Hackback, Stiftung Neue Verantwortung and Dorothy E. Denning and Bradley J. Strawser (2017), Active Cyber Defense: Applying Air Defense to the Cyber Domain, Georgetown University Press and HM Government (2016), National Cyber Security Strategy 2016-2021, United Kingdom HM Government and The Cybersecurity Tech Accord (2020), No Hacking Back: Vigilante Justice vs. Good Security Online, The Cybersecurity Tech Accord

37 Dorothy E. Denning and Bradley J. Strawser (2017), Active Cyber Defense: Applying Air Defense to the Cyber Domain, Georgetown University Press

38 Jack Goldsmith and Alex Loomis (2021), "Defend Forward" and Sovereignty, Hoover Institution

Second, active cyber defense is generally considered a response to a specific[39] *malicious cyber operation or campaign*. Although debates on preemptive applications of active cyber defense, mainly surrounding *defend forward* and *persistent engagement*,[40] have surfaced, active cyber defense is considered not to be preemptive per se. Of course, when active cyber defense is a response to a malicious cyber campaign,[41] the measure can be executed between operations of the same campaign, thus preempting operations that have not yet been executed. An example is the takedown of a botnet command-and-control structure.[42] By the time the command-and-control structure is taken down, it may have been leveraged in malicious activities, making the takedown neither entirely *ex ante* nor entirely *ex post*.

Third, compared to passive forms of cyber defense, such as firewalls and intrusion detection systems, active cyber defense does not primarily improve IT security. The objective for any active cyber defense operation is to *technically neutralize and/or mitigate the impact of a malicious cyber operation or campaign and/or attribute it technically*. The technical attribution can then lead to the implementation of additional response (political) mechanisms such as sanctions.[43] Therefore, active cyber defense operations generally do not aim for retribution but to end or at least decrease the effects of the malicious cyber activity.

This paper focuses on *government-led* active cyber defense: activities carried out solely by government entities or *led or coordinated by government entities* with support from the industry and possibly other sectors. IT companies and the IT security industry can play an important role in supporting active cyber defense operations,[44] especially for complementary attribution insights[45].

Active cyber defense operations are not limited to protecting a given country's government, military, academia, industry including critical infrastructure and society, as (under observation of international law, especially regarding countermeasures[46])

39 Dorothy E. Denning and Bradley J. Strawser (2017), Active Cyber Defense: Applying Air Defense to the Cyber Domain, Georgetown University Press

40 Jacquelyn G. Schneider (2019), Persistent Engagement: Foundation, Evolution and Evaluation of a Strategy, LAWFARE

41 Regarding the difference between *cyber operation and cyber campaign*, see Richard J. Harknett and Max Smeets (2020), Cyber campaigns and strategic outcomes, Journal of Strategic Studies

42 Dorothy E. Denning and Bradley J. Strawser (2017), Active Cyber Defense: Applying Air Defense to the Cyber Domain, Georgetown University Press

43 Kristen E. Eichensehr (2019), Decentralized Cyberattack Attribution, American Journal of International Law Unbound and Sven Herpig (2021), Die Beantwortung von staatlich verantworteten Cyberoperationen, Konrad-Adenauer-Stiftung

44 For example, The Cybersecurity Tech Accord (2020), No Hacking Back: Vigilante Justice vs. Good Security Online, The Cybersecurity Tech Accord and Sven Herpig und Dennis-Kenji Kipker (2021), Der Zweck heiligt nicht die Mittel, Netzpolitik.org

45 Kristen E. Eichensehr (2019), Decentralized Cyberattack Attribution, American Journal of International Law Unbound

46 United Nations (2005), Responsibility of States for Internationally Wrongful Acts, United Nations and Ashley Deeks (2020), Defend Forward and Cyber Countermeasures, Hoover Institution

they may be extended to allied countries as a means for *collective* active cyber defense. This includes, but is not limited to, joint law enforcement operations—reflecting the potential application of active cyber defense across sectors against crime, intelligence operations and military measures alike.

Selected examples of measures THAT FALL under the definition of active cyber defense:

- Mandating Internet Service Providers (ISPs) block or reroute malicious traffic;
- Mandating ISPs lock compromised customer systems in a walled garden/sandbox, displaying information on how to clean up and patch the systems;
- Setting up a sinkhole to take over the command-and-control infrastructure used in malicious cyber campaigns, for example, of botnets;
- Deploying or mandating the deployment of beacons within their own perimeter (as part of a honeypot)[47] that, when copied and executed, signal their current location, or take other actions on the target;
- Mandating ISPs deliver updates and respective notifications to their customers for software and hardware beyond devices provided by the ISP;
- Taking over a command-and-control infrastructure used in malicious cyber campaigns to uninstall or neutralize malware on the victims' systems and/or deploy patches;
- Exploiting vulnerabilities and deploying malware to compromise the IT infrastructure of perpetrators to monitor their activities, technically attribute a malicious cyber campaign or disrupt their activities.

Selected examples of measures THAT DO NOT fall under the definition of active cyber defense:

- Deploying firewalls, malware detection software, intrusion detection systems and similar software;
- Sharing indicators of compromise and other data with security vendors to adjust or reconfigure their products;
- Setting up or mandating the setup of honeypots to gather intelligence on a malicious cyber operation or campaign;
- Compromising IT infrastructure to disrupt information operations (e.g., Operation GLOWING SYMPHONY[48]).

47 Gregory Falco and Herb Lin (2018), Active Cyber Defense and Interpreting the Computer Fraud and Abuse Act, LAWFARE

48 National Security Archive (2020), USCYBERCOM After Action Assessments of Operation GLOWING SYMPHONY, National Security Archive

It is clear that the definition is still very broad and includes many less controversial activities, such as government-led sinkholing malicious traffic or mandating that ISPs block malicious traffic. This broadness is intentional to prevent a binary debate about whether to allow or disallow active cyber defense per se. The idea is to present a common framework for assessing the associated risks and opportunities of individual measures. However, the focus of the debate will likely be controversial activities such as exploiting vulnerabilities in the IT systems of the groups behind a malicious operation or campaign to disrupt their infrastructure. Although it is challenging to separate more controversial from less controversial active cyber defense measures in general, the criteria discussed in the following section can assist in evaluating the level of controversy for a particular measure.

3. Criteria

As states vary in their security culture and policies, legal frameworks and cyber security strategies, as well as in their resources, capacities and capabilities, it is not prudent to decide whether an active cyber defense operation would be considered lawful merely based on the definition.

When calculating the risks for (unintended or cyber-physical, especially in critical infrastructure) damage, fundamental rights violations, violations of sovereignty, conflict escalation and possible success, there is not a common measurement for active cyber defense. Therefore, very loosely based on the discussions on vulnerability equity processes (VEPs)[49] and government disclosure decision processes (GDDPs)[50], the different criteria should be examined and assessed before deploying a tool or using a measure. That includes, but is not limited to, the *who*, *against whom*, *where*, *to what effect* and *when* of an active cyber defense operation.

Each criterion with its different indicators enhances the approximation of the implications of the specific measure. However, the different indicators cannot be hard coded with risk levels. Analyzing the implications depends on additional circumstances. For example, an active cyber defense operation carried out by a law enforcement agency is, operationally speaking, as risky as the same operation carried out by the military. However, legal and cultural aspects might kick in when, for example, the operation takes place domestically, and the deployment of military capabilities at home is highly problematic.

The final decision to conduct an active cyber defense operation (and whether it is a responsible decision[51]) rests upon the summarized implications across the individual criteria assessed in accordance with the implemented safeguards. Therefore, decisions about deployment of active cyber defense operations must always be made on a case-by-case basis.

49 Rob Joyce (2017), Improving and Making the Vulnerability Equities Process Transparent is the Right Thing to Do, The White House

50 For example, Sven Herpig and Ari Schwartz (2019), The Future of Vulnerabilities Equities Processes Around the World, LAWFARE

51 Compare to the debate on responsible cyber offense; see Perri Adams, Dave Aitel, George Perkovich and JD Work (2021), Responsible Cyber Offense, LAWFARE

Category	Criterion	Indicators
Purpose	Goal	Mitigation
		Neutralization
		Attribution
	Success	Strategic
		Tactical
		None
Effect	Type	Non-intrusive
		Reversible
		Non-reversible
		Intrusive
	Space	Blue space
		Green space
		Red space
		Gray space
	Target	Non-critical infrastructure
		Critical infrastructure
Actors	Government lead agency	Cybersecurity agency
		Law enforcement
		Intelligence agency
		Military
	Cooperativeness	Cooperative
		Unknown
		Noncooperative
Timing	Attribution	Not required, not necessary
		Uncertain
		Fairly Certain
		Very Certain
		Proven

Category	Criterion	Indicators
	Time	During operations of the same campaign
		In-between sequential campaigns
		Immediately after an operation
		Some time after an operation
Operations	De- and escalation	Potential de-escalation
		No change in the escalation cycle
		Potential escalation
		Proportional
		Not proportional
	Automation	Automated
		Semi-automated
		Manual
	Frequency	One-off
		Periodic
		Sustained
	Cost	Low
		High
	Collateral consequences	Not expected
		Known unknowns
		Expected

Table 1.
Criteria for active cyber defense operations

3.1. Purpose

3.1.1. Goal

The goals of an active cyber defense operation are derived directly from its definition. An operation can aim to mitigate the impact, neutralize a malicious cyber operation or campaign and/or attribute it technically. The last may serve as a foundation for follow-up policy measures to responses[52] as well. To increase security immediately, neutralizing a malicious cyber campaign and/or mitigating its impact should be prioritized over attributing it technically, if not required for either. Goals such as deterrence by punishment or retribution are not considered, as there is no evidence that these aspects contribute directly to defending against an ongoing malicious cyber activity. Ideally, however, active defense would, over time and as a second-order effect, serve to deter malicious activity in general by imposing costs on the threat actor. However, the focus of any specific active cyber defense operation should always be to mitigate or neutralize the immediate threat. Additionally, a cyber operation carried out for the sole purpose of punishment or retribution may increase the risk of escalation and may violate international law[53]. Operations aiming for those goals would fall under the category of offensive (military) cyber operations[54] instead.

3.1.2. Success

Most of the criteria described above are risk-related. However, on the other side of the equation stands the effect or, more accurately, the envisioned effect. To better estimate the success of the effect, it is crucial to look at the expected goal. Of course, assessing this criterion for any given active cyber defense operation works much better *ex post* but is of only little help to an *ex ante* impact assessment. However, it should not be discarded. A possible categorization, connected to the frequency of operations, is whether the operation is likely to result in a *tactical* or *strategic* success. If, for example, the takedown of a botnet led only to quick relief before other operators exploit the vacuum left, it may be not too difficult to assign a botnet-takedown operation the label “tactical”. Collecting the last pieces of digital evidence for the attribution of a long-lasting malicious cyber campaign that will be used as the basis for (cross-domain) countermeasures could be labeled “strategic”. In both cases, active cyber defense operations could be labeled during the *ex ante* impact assessment.

52 Sven Herpig (2021), Die Beantwortung von staatlich verantworteten Cyberoperationen, Konrad-Adenauer-Stiftung
53 Michael Schmitt (2021), Three International Law Rules for Responding Effectively to Hostile Cyber Operations, JUST SECURITY
54 Matthias Schulze (2020), Militärische Cyber-Operationen, Stiftung Wissenschaft und Politik

Concerning the overall assessment of active cyber defense operations, a higher-risk operation, therefore, should lead to a strategic success rather than just to a tactical success.

3.2. Effect

3.2.1. Type

A prime criterion that should be considered is the desired effect that the planned active cyber defense operation will have, and an acknowledgment that the achievable effect may differ from the intended. The effects might range widely in terms of intrusiveness, which is one of the key aspects to consider when looking at this criterion.

Closely connected to this aspect and to the limitations of countermeasures under international law is the potential "reversibility" of measures.[55] This is also crucial from the technical perspective. If an active cyber defense operation leads to unintended consequences, reversibility ensures damage control.

Thus, although there are non-intrusive, reversible effects, there are also intrusive, non-reversible measures. Whereas the intrusiveness of measures is not necessarily binary but on a spectrum, it may be useful to divide them into *intrusive* and *non-intrusive* methods for operationalizing the framework. For example, the former includes measures that meddle with the confidentiality, integrity or availability of IT systems, infrastructure and data.

3.2.2. Space

Closely linked to the type of effect is the space of the effect. Active cyber defense operations can be categorized into different areas of effect[56] or effect spaces. For the purpose of this paper, the spaces are connected directly to the states in which the IT systems that are targeted with active cyber defense measures are located independently of whether the IT systems are part of the threat actor's internal infrastructure (e.g., the devices they are operating from) or innocent third parties (e.g., bots). This perspective is in line with the concept of due diligence and geopolitical risks for the implementer of active cyber defense operations. The distinctions between these spaces vary across concepts. This paper differentiates *blue space*, *green space, red space* and *gray space*.

55 Ashley Deeks (2020), Defend Forward and Cyber Countermeasures, Hoover Institution
56 For example, Max Smeets (2019), Cyber Command's Strategy Risks Friction With Allies, LAWFARE and Jason Healey, Neil Jenkins and JD Work (2020), Defenders Disrupting Adversaries: Framework, Dataset, and Case Studies of Disruptive Counter-Cyber Operations, CyCon 2020

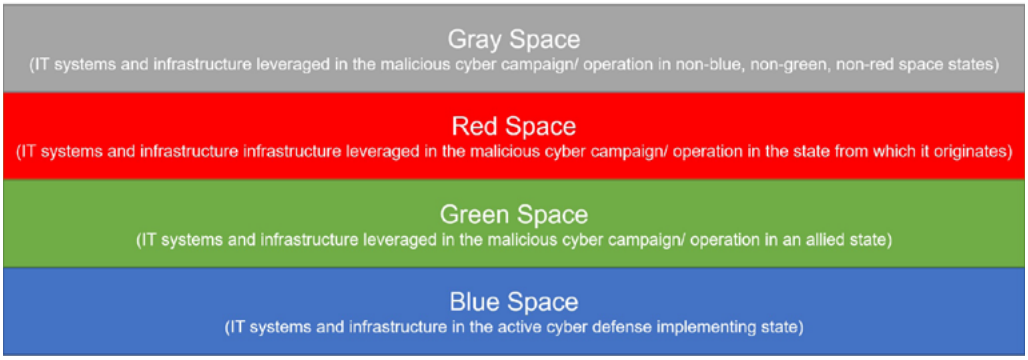

Figure 1.
Effect spaces for active cyber defense operations

The definition of active cyber defense in this paper requests government leadership in the operations. Therefore, *blue space* is defined as the area within the jurisdiction of the government, including the private sector among others[57]. Blue space, therefore, entails all IT systems and infrastructure of a government, the country's inhabitants and companies located in it that are affected in the malicious cyber activity that the government wants to defend against.

Green space[58] is defined as IT systems and infrastructure affected in a malicious cyber operation or campaign that are within the jurisdiction of an allied government. Although there is always the risk of third-party collection through an ally, existing and future treaties, joint task forces, communication channels and other instruments enable various cooperative ways to counter malicious cyber activities. These options must be considered seriously before unilateral action is taken against the IT systems and infrastructure of an allied government, even when they are (unknowingly) involved in a malicious cyber activity. Operations carried out in green space may also have less potential for backfiring if an ally benefits from them. The FBI's Operation Torpedo (although government hacking and not an active cyber defense operation) and the BKA's disabling of Emotet malware resulted in manipulation of integrity and intrusive access in computer systems across the globe.[59] However, none of the affected actors' governments complained publicly about this manipulation. From an international law perspective, the only feasible argument for a unilateral, noncooperative active cyber defense operation in allied jurisdictions would be their failure to comply with due diligence obligations.

Red space is defined in this paper as IT systems and infrastructure used in a malicious cyber activity that are within the jurisdiction of the country in which the operation or campaign originates. Although there is a risk of conflict escalation, active

57 For example, Justin Hendy (2020), Govt introduces cyber incident response takeover bill to parliament, nextmedia
58 In this paper, different from existing definitions, the concept of *gray space* is divided into *green space and gray space*.
59 Joseph Cox (2016), The FBI's 'Unprecedented' Hacking Campaign Targeted Over a Thousand Computers, VICE and Andre Meister (2021), BKA nutzt Emotet-Takedown als Türöffner für mehr Befugnisse und neue Gesetze, Netzpolitik.org

cyber defense operations in red space may be more in line with international law requirements (e.g., for countermeasures[60]) than in green or gray space. Of course, there are different levels of national responsibility for malicious cyber operations or campaigns that are carried out from a country's territory. On Healey's "Spectrum of State Responsibility,"[61] *state-prohibited* and *state-prohibited-but-inadequate* could lead to considering the space as *green* or *gray space* (depending on diplomatic relations) rather than *red space* and treating it as such.

Gray space is defined in this paper as IT systems and infrastructure used in a malicious cyber operation or campaign that are not located in blue, green or red space. Measures that can be taken are much more limited than in the other spaces and possibly more escalatory. From an international law perspective, the only feasible argument for a noncooperative active cyber defense operation in green and gray spaces may be their failure to comply with due diligence obligations. Therefore, from a subsidiarity point of view, a malicious cyber operation or campaign should be defended in gray space only if all other possible measures in blue space, green space (if cooperative) and red space have been exhausted.

The categorization assumes correct, confident technical attribution.[62] Thus, the assessment may change over the course of planning and implementing an active cyber defense operation (e.g., when a false flag operation is encountered) and, therefore, adapt to the new information. At the same time, due to the distribution of bots in a botnet or division of labor among various actors in a cybercrime campaign,[63] an active cyber defense operation may target IT systems in a number of different spaces with different types of effects. Thus, the space criterion is abstract and may be challenging to operationalize.

3.2.3. Target

Irrespective of their location within the blue, green, red or gray space, concrete target systems and infrastructure must be considered and possibly treated differently, especially if the type of effect is intrusive and non-reversible. Although there are, of course, differences between government infrastructure, industry systems and personal devices, it is crucial to make a clear distinction between critical infrastructure and non-critical infrastructure. There is some debate about what counts as a critical infrastructure,[64] but it is a solid rule of thumb to treat targets of active cyber defense operations as critical infrastructure if they would be considered critical infrastruc-

60 Ashley Deeks (2020), Defend Forward and Cyber Countermeasures, Hoover Institution
61 Jason Healey (2011), Beyond Attribution: Seeking National Responsibility for Cyber Attacks, Atlantic Council
62 Aleksandra Sowa (2016), Der Kalte Krieg ist vorbei – es lebe der Kalte Cyber-Krieg!, The European
63 Bundeskriminalamt (2021), Bundeslagebild Cybercrime 2020, Bundeskriminalamt
64 The Federal Government of Germany (2021), On the Application of International Law in Cyberspace, German Federal Foreign Office

ture in the nation conducting active defense. Needless to say, targeting critical infrastructure with active cyber defense operations should be assessed with utmost care, as it can lead to unintended and even cyber-physical effects and the subsequent risk of escalation.

3.3. Actors

3.3.1. Government lead agency

When discussing the implementer of active cyber defense operations, several types of government actors are considered for the lead agency. They include national cybersecurity agencies[65], federal and state law enforcement agencies[66], intelligence agencies[67] and the military[68]. Although centralizing active cyber defense operations has a certain appeal, it may be hard due to the plethora of different measures covered by this definition.

Considering the effect space, cybersecurity and law enforcement agencies may be the appropriate lead for measures in blue and green spaces and intelligence agencies for measures in gray and/or red spaces. However, there are many more aspects to consider, such as capabilities, availability of intelligence information or access to classified information, existing cooperation with involved national and international actors, political will of the agency heads, foreign policy implications etc. Countries also have varying strategic cultures, and their perspectives on the respective agencies naturally differ. Furthermore, the military carrying out an active cyber defense operation may present a different signaling and escalatory effect than a law enforcement agency doing it.

Ultimately, different stakeholders may be directly involved in an active cyber defense operation. To coordinate these measures, especially where the lines get blurry, it would be prudent to discuss all assumed active cyber defense operations on a joint operational platform, preferably one that already exists. Keeping that in mind, there must be a central focal point at the strategic level overseeing all active cyber defense operations and responsible for adjusting the overfall framework. For the United States, this could be, for example, the national cyber director; for Germany, this position does not exist yet[69].

65 For example, the Federal Office for Information Security in Germany or the Cybersecurity and Infrastructure Agency in the United States
66 For example, the Federal Criminal Policy office in Germany or the Federal Bureau of Investigation in the United States
67 For example, the Federal Foreign Intelligence Service in Germany or the National Security Agency in the United States
68 For example, the German Cyber and Information Space Command or the U.S. Cyber Command
69 Sven Herpig (2021), Die Beantwortung von staatlich verantworteten Cyberoperationen, Konrad-Adenauer-Stiftung

3.3.2. Cooperativeness

Independent from effect space, although more likely to be found in blue space, cooperativeness should be considered: whether the parties involved in intermediary steps or at the receiving end of active cyber defense operations (owners of the IT systems and their respective governments) cooperate with the implementing government agency and whether that is known by the government lead agency. According to Denning and Strawser, "[d]efenses become noncooperative when they involve actions taken against external computers without the permission of the user or network owner."[70]

Cooperation is important, as, for example, the risks and resources involved in exchanging information with an affected third party unwittingly involved in a malicious cyber operation or campaign may be lower than attempting to compromise that third party through an active cyber defense operation to acquire said information against the party's will. Risks are still involved even in this example, as cooperation may trigger unwanted third-party collection, especially in gray spaces, that could have otherwise been avoided.

Moreover, just because an active cyber defense operation, for example, neutralizes malware on companies' or citizens' IT systems (compare, for example, the removal of the Hafnium web shell and the takedown of the Emotet botnet), it cannot, although likely, simply be assumed that the operation is cooperative as long as the target has not consented to it. Although urgency or other operational concerns may hinder *ex ante* cooperation with actors, it may reduce risks, for example, of escalation.

3.4. Timing

3.4.1. Attribution

Several active cyber defense measures do not require previous technical attribution of the malicious cyber operation or campaign they are implemented against, and/or are even used to increase confidence in the technical attribution, such as rerouting malicious traffic or beacons. For all other measures, attribution through technical intelligence and other means plays a major role. Although active cyber defense measures can contribute to technical attribution, implementing (intelligence) activities to attribute cyber operations and campaigns (reconnaissance) may be required before certain active cyber defense measures can actually be implemented. Finlay

70 Dorothy E. Denning and Bradley J. Strawser (2017), Active Cyber Defense: Applying Air Defense to the Cyber Domain, Georgetown University Press

and Payne, for example, explained that "[a] state often cannot practically respond to a threat unless it knows from where the threat emanates and potentially who is responsible."[71] Additionally, countermeasures, which may be active cyber defense operations, justified by a violation of sovereignty under international law (internationally wrongful acts) require a certain level of attribution regarding the origin of the malicious cyber operation or campaign.[72] Different agencies may have different confidence levels for technical attribution. Whatever they are, they are crucial as the basis for a number of active cyber defense measures—both technically and legally.[73] After considering all the technical, intelligence and (geo)political evidence[74], assigned confidence levels may be, for example, *uncertain, fairly uncertain, very certain* or *proven*.

3.4.2. Time

A component closely linked to the confidence level of attribution is *time*. In general, confidence in attribution benefits from more time for those conducting the analysis. However, the more time passes between the incident(s) and the active cyber defense operation, the less effective the response may be, for example, in terms of damage mitigated.[75] If an active cyber defense operation that is designed as a countermeasure is, in terms of time, too far disconnected from the malicious cyber operation or campaign the active cyber defense operation seeks to disrupt, it may fail to meet the necessary self-defense criteria (e.g., ending that activity or securing reparations[76]). Therefore, the operation may be seen as pure retribution and considered an internationally wrongful act.

Similarly, a challenge arises when active cyber defense operations are conducted before a malicious cyber operation or campaign is launched. Although that may be admissible for persistent engagement and defend forward[77], it is, as described in the definition, not applicable to active cyber defense operations. Exceptions are tools that are implemented as preventive measures and triggered during an operation or campaign (e.g., beacons).

71 Lorraine Finlay and Christian Payne (2019), The Attribution Problem and Cyber Armed Attacks, American Journal of International Law Unbound
72 Michael Schmitt (2021), Three International Law Rules for Responding Effectively to Hostile Cyber Operations, JUST SECURITY
73 For example, Kristen E. Eichensehr (2019), Decentralized Cyberattack Attribution, American Journal of International Law Unbound and Ashley Deeks (2020), Defend Forward and Cyber Countermeasures, Hoover Institution
74 Sven Herpig and Thomas Reinhold (2018), Spotting the bear: credible attribution and Russian operations in cyberspace, European Union Institute for Security Studies
75 Thomas Reinhold und Matthias Schulze (2017), Digitale Gegenangriffe, Stiftung Wissenschaft und Politik
76 Michael Schmitt (2021), Three International Law Rules for Responding Effectively to Hostile Cyber Operations, JUST SECURITY
77 Jacquelyn G. Schneider (2019), Persistent Engagement: Foundation, Evolution and Evaluation of a Strategy, LAWFARE

Therefore, the window of opportunity for conducting active cyber defense operations may be narrow. However, as active cyber defense is not limited to countering individual operations but can also be used as a response against campaigns, it can take place during operations of the same campaign or in-between sequential campaigns. Depending on the kind of active cyber defense operation, preparations may take months or longer.[78] Timing-wise, and considering legal aspects, preparation, attributability and risks, finding the sweet spot may be challenging.

3.5. Operations

3.5.1. De- and escalation

Several active cyber defense measures carry the potential risk for escalation, although they are likely to remain within the cyber domain[79]. At the same time, not responding to an ongoing malicious cyber operation or campaign may also lead to an escalation in a different way, as it potentially encourages threat actors to continue or even grow bolder with their activities.

The potential for de- and escalation is not only limited to the government leading the active cyber defense efforts and the threat actor and its respective government but also includes third parties that need to be factored in, especially when operating in green, gray and red spaces. Spillover effects, second- and third-order effects and unintended consequences are difficult to calculate and partially also based on the controllability and quality of the active cyber defense measures deployed.

However, it does not follow that a state implementing active cyber defense measures to defend itself from a malicious operation or campaign will always trigger an escalatory spiral. The fact that thus far a state's offensive cyber operations may have not appeared to spiral out of control[80] yet may support that point.

Adhering to proportionality, however, is prudent to maintain legality and avoid escalation if possible. As proportionality is also recognized as a requirement for countermeasures under international law[81], active cyber defense operations must strive to be as proportional as possible.

78 Lukas Mäder and Georg Häsler (2021), Interview "Ein Cyberangriff der Armee würde Monate oder Jahre dauern", Neue Züricher Zeitung
79 Matthias Schulze, Josephine Kerscher and Paul Bochtler (2020), Cyber Escalation, Stiftung Wissenschaft und Politik
80 Jason Healey and Robert Jervis (2020), The Escalation Inversion And Other Oddities Of Situational Cyber Stability, Texas National Security Review and Jason Healey (2019), The implications of persistent (and permanent) engagement in cyberspace, Journal of Cybersecurity
81 Ashley Deeks (2020), Defend Forward and Cyber Countermeasures, Hoover Institution

Additionally, the current geopolitical context of the parties[82] and the nature of the malicious cyber operation or campaign and the stakeholders (criminals, intelligence agencies, military) behind it are crucial to estimate the de- and escalation potential.

Confidence-building measures such as open communication channels can play a vital role in avoiding escalation and lead to de-escalation if desired.

3.5.2. Automation

Linked to the question of escalation is the level of automation—especially in the sense of speed, scale and control (e.g., human-in-the-loop or -on-the-loop)—of an active cyber defense operation. Take, for example, law enforcement measures that roll out patches to all infected IT systems of a botnet after taking over its command-and-control infrastructure (automated deployment of software) or nematode software that scans for vulnerable Internet of Things devices and patches them automatically (self-propagation and intrusion). These are two very different examples of automation and associated risk levels.

Increased automation when targeting heterogeneous systems and infrastructure may lead to unintended consequences, as loss of control can lead to disastrous results.[83] Therefore, a higher standard of care is possible if measures are implemented individually and manually. However, manual changes are also prone to errors. Moreover, such an approach may not be feasible due to the sheer number of targeted systems; the takedown of the Emotet botnet likely affected tens of thousands of computers[84]. Additionally, there may not be sufficient time to adequately develop the required software, leaving it with vulnerabilities or unexpected effects on the target, for example, through a lack of testing in specific environments.

Automated or semi-automated measures can and should be safeguarded, for example, through kill switches.[85] However, manual measures should also be safeguarded, for example, through the dual control principle.

How risky an automated process is depends on other criteria, such as the type of effect. Although signaling beacons triggering automatically and indiscriminately may be a desired function, the automatic and indiscriminate triggering of trap files that encrypt the adversaries' devices may not always be advantageous in the overall picture, for example, factoring in the escalation criterion.

82 Jason Healey and Robert Jervis (2020), The Escalation Inversion And Other Oddities Of Situational Cyber Stability, Texas National Security Review and Thomas Reinhold und Matthias Schulze (2017), Digitale Gegenangriffe, Stiftung Wissenschaft und Politik
83 Ciaran Martin (2020), Cyber-weapons are called viruses for a reason: Statecraft and security in the digital age, King's College London
84 Anna Biselli (2021), Darf das BKA Schadsoftware auf infizierten Rechnern manipulieren?, Netzpolitik.org
85 Perri Adams, Dave Aitel, George Perkovich and JD Work (2021), Responsible Cyber Offense, LAWFARE

3.5.3. Frequency

The overall calculation of whether an active cyber defense operation makes sense should factor in the frequency of measures needed to achieve the goal. In terms of efficiency and (political) effectiveness, a one-off operation may be preferable to an operation that needs to be repeated regularly to achieve the goal. If repetitions of active cyber defense operations are necessary, and they often may be as threat actors react and adapt their infrastructure after a takedown[86], other measures such as passive cyber defense or a combination of both may become more effective and efficient. However, regular practice may increase the capabilities of the actors implementing active cyber defense operations through sustained practice and learning.

Although the removal of the Hafnium web shell was a one-off operation, the IT systems were left unpatched, and threat actors could have compromised the systems again in the same way and left another web shell, technically requiring another operation to remove the web shells. While providing some methodological caveat concerning their dataset, Healey, Jenkins and Work stated that

> *"[c]ommonalities across the entirety of the case dataset importantly suggest that operational disruption is rarely accomplished as a single decisive action, at least where adversary operators, developers, and planners continue to enjoy a sustained base of uninterrupted support. However, merely because a single action will not render the adversary hors de combat does not negate the utility of disruption. Forcing adversary adaptation may add value, particularly where such a response requires investment disproportionate to the value of continuing operations or where adversary resourcing may be constrained in some other dimensions."*[87]

The authors offer a useful three-point scale of *one-off*, *periodic* and *sustained*.[88]

3.5.4. Costs

Closely linked to the frequency criterion are the costs for an active cyber defense operation in terms of resources. If, for instance, an ISP has to be mandated to block certain malicious traffic regularly, that may not lead to huge costs because it needs to be implemented once and then operates automatically, although cost reimburse-

86 Rob Joyce (2021), Risky Biz Feature Podcast: An interview with Rob Joyce, Risky Business and Eric Rosenbach, Juliette Kayyem and Lara Mitra (2021), The Limits of Cyberoffense, Foreign Affairs

87 Jason Healey, Neil Jenkins and JD Work (2020), Defenders Disrupting Adversaries: Framework, Dataset, and Case Studies of Disruptive Counter-Cyber Operations, CyCon 2020

88 Jason Healey, Neil Jenkins and JD Work (2020), Defenders Disrupting Adversaries: Framework, Dataset, and Case Studies of Disruptive Counter-Cyber Operations, CyCon 2020

ment must be factored in[89]. In contrast, an active cyber defense operation that targets the threat actor's IT systems with the goal of wiping all their data may be a one-off. However, the cost of specialized-staff time, procurement of tools and exploits and third-party support from the private sector may make that operation prohibitively expensive, or at least inefficient. Together with other criteria, this criterion may render the entire operation not only inefficient but also ineffective in terms of achieving the overall goal. However, the damage and costs inflicted by the malicious cyber operation or campaign if left unanswered must also be considered.

3.5.5. Collateral consequences

Although the *type*, *space*, *attribution* and *de-* and *escalation* criteria already touch on the topic of collateral consequences, it is prudent to spell them out explicitly. Sasha Romanosky and Zachary Goldman pointed out that "cyber collateral damage" deserves special attention, arguing that

> *"given the interconnectedness of cyber and cyber-physical systems, direct, indirect, and collateral effects can be much more difficult to predict, rendering ineffective traditional approaches to collateral damage estimation (CDE)."*[90]

Collateral consequences go beyond simply referring to accidental consequences, as actions may be taken while it is clear that collateral consequences could occur. Planners of active cyber defense operations may either expect (and accept or not accept) or not expect collateral consequences. Moreover, they may be right or wrong about their analysis. However, at this stage, it is also useful to point out known unknowns and interdependencies. For example, the software versions of the targeted systems of an active cyber defense operation can be known unknowns that may or may not lead to malfunctioning of the malware-removal software that is supposed to be deployed and was tested only on systems running the latest software version.

89 Compare, for example, Ryan Guthrie (2020), CALEA Compliance and Cost Reimbursement, Advanced Technologies and Services

90 Sasha Romanosky and Zachary Goldman (2017), Understanding Cyber Collateral Damage, Journal of National Security Law & Policy

4. Safeguards

Using the definition provided in this paper, it can be estimated whether a planned deployment of measures constitutes an active cyber defense operation. The criteria and their indicators offer an analytical framework for examining the operation's crucial elements to better assess the operation's risks, usefulness and potential costs. Although these steps are aimed at a case-by-case analysis, governments deciding to implement measures from the broad range that the definition offers also need to implement structural and procedural safeguards and apply them to all active cyber defense operations. These structural safeguards will *inter alia* guarantee privacy protections, ensure alignment with human rights and national and international law, maintain geopolitical stability and create a net-gain for national security.

Safeguard	Subsumed safeguards
Define and limit the scope	
Establish a national legal framework	Require impact assessments Implement oversight Create transparency and auditability
Set up guidelines for tools and services	
Apply international law	
Consider public interest	
Adapt confidence-building measures	

Table 2. Safeguards for active cyber defense operations

4.1. Define and limit the scope

One safeguard to be clearly outlined in the warrant and impact assessment is the scope of the operation. It should be as narrow as possible and at least assess whether the targets of an operation are in blue, green, gray or red space. Ideally, the scope is limited to blue space, as this normally means directly supporting the targets of a malicious cyber operation or campaign within one's own jurisdiction, rather than going after the threat actor. An example is the *removal of the Hafnium web shells*.

Noncooperative operations in green and gray spaces as well as operations in red space in general carry high risks and may lead to unintended consequences. Therefore, noncooperative operations in these spaces should be regarded as the *ultima ratio*. Talking about a subset of these measures, the authors of *Into the Gray Zone* argued that

> *"[f]rom a policy standpoint, these techniques are likely to escalate incidents because they are likely to be disproportionate, difficult to adequately contain, often retributive, and imprecise. Activities such as hacking back to retrieve stolen data or infecting the attacker's own systems with malware are likely to be ineffective. Stolen data likely exists in a multitude of locations both inside and outside of the attacker's networks. It is unlikely that an attempt to retrieve stolen data would remove it from every location where it is stored; the risk of escalation in exchange for uncertain gains is simply too high. These types of cyber defense activities should continue to be prohibited."*[91]

Ciaran Martin, former head of the United Kingdom's National Cyber Security Centre, stated that

> *"'hacking back' will not 'deter' cyber espionage, which is generally accepted under international norms. And on the relatively rare occasions when those norms are crossed, the sorts of capabilities offensive cyber affords are generally not appropriate ones for pushback."*[92]

4.2. Establish a national legal framework

Next to international regulation, creating a national legal framework for active cyber defense operations is crucial for several reasons, especially due to the privacy invasiveness and security risks that may be inherent to active cyber defense operations, and because it is vital to clarify the rules applying to everyone involved. Adherence to the rule of law may be facilitated by having a clear and possibly specific legal framework for active cyber defense operations. Additionally, a legal framework can and should include a number of safeguards and clear assignments of liability that apply to active cyber defense operations.

Only then can government agencies be certain that their measures are lawful and that citizens and other parties are appropriately protected from the state. Deriving investigatory and other powers from non-cyber-specific laws should be the exception, not the norm.

The legal framework and its elements should be regularly revisited, evaluated and refined in accordance with national sunset clause procedures.

91 Center for Cyber and Homeland Security (2016), Into the Gray Zone - The Private Sector and Active Defense Against Cyber Threats, The George Washington University
92 Ciaran Martin (2021), Offensive cyber in the age of ransomware, Offensive Cyber Working Group

The positive effect of having a legal framework can be seen in a comparison of the aftermath of the removal of the Hafnium web shells and the disabling of Emotet malware. Debates about and critics of the former did not question the legality of the operation[93]; however, for the latter, the legality was the center of attention[94]. The last thing a government wants after a successful operation (and wants even less after an unsuccessful one) is (legitimate) doubt about the operation's lawfulness.

4.2.1. Require impact assessments

A formal *ex ante* impact assessment is essential to weigh the risks, impact, chances and possible consequences of an operation, develop additional options and back-up plans and define circuit breaker conditions. The criteria discussed above play a major role when drafting the impact assessment. It should also speak to why the envisioned active cyber defense operation is the least intrusive and most promising option on the table. Therefore, the impact assessment should also include a discussion of the impact of not taking the action. Ideally, the impact assessment includes independent technical, legal, economic and policy expertise.

Nevertheless, it is clear that any *ex ante* impact assessment often has to work with limited information and, therefore, provides only a narrow picture. In the case of so-called "black box" assessments, where little information is available in advance, a set of predefined general criteria is suitable to perform a self-assessment. Additionally, adaptations may occur during the operation. However, the impact assessment is a useful basis for decision-makers and oversight bodies and, therefore, should be a requirement for every active cyber defense operation.

Impact assessments must be part of the legal framework.

4.2.2. Implement oversight

Active defense operations most likely have extraterritorial implications, touching on sovereign issues of other states. Therefore, a high standard of oversight is required and should be enshrined in the legal framework.

93 For example, Catalin Cimpanu (2021), FBI operation removed web shells from hacked Exchange servers across the US, The Record and Alex Iftimie (2021), No Server Left Behind: The Justice Department's Novel Law Enforcement Operation to Protect Victims, LAWFARE and Ed Amoroso and Randal S. Milch (2021), Hack-to-Patch by Law Enforcement Is a Dangerous Practice, JUST SECURITY

94 Dennis-Kenji Kipker and Michael Walkusz (2021), Das BKA zerschlägt die Infrastruktur von Emotet: Mit welcher Rechtsgrundlage?, beck-community and Sven Herpig and Dennis-Kenji Kipker (2021), German Emotet takedown in the legal gray zone, Stiftung Neue Verantwortung and Andre Meister (2021), BKA nutzt Emotet-Takedown als Türöffner für mehr Befugnisse und neue Gesetze

All active cyber defense operations led by law enforcement should require an *ex ante* warrant, including specific parameters of the planned operation, for example, in the form of an impact assessment. Although active cyber defense operations take time to prepare, there may be edge cases in which imminent danger can be responded to quickly by an active cyber defense operation. In those cases, immediate measures followed by a mandatory retroactive judicial review should be possible. Specialized courts should be in charge and provided with technical capacity-building and the option to bring in independent technical expertise to enable judges to understand the possible implications of the operations stated in the warrant.

All active cyber defense operations in green (if noncooperative), gray and red spaces should require approval from the highest echelons of government, such as the prime minister, chancellor or the secretary/minister of justice, homeland security or defense, as the consequences of the operations may go well beyond the area of responsibility of the implementing agency.

Additionally, after-action reports of all active cyber defense operations must be provided to a parliamentary oversight committee for additional scrutiny, where feasible. After-action reports should start with the impact assessment and include the achieved goals as well as the observed intended and unintended negative consequences and possible mid- and long-term implications.

Last, to facilitate remedial actions, all active cyber defense operations should include a notification of the targets in blue space. The notification needs to include the particularity requirements laid out in the warrant application if a warrant was involved. Due to operational concerns, the notice does not have to be provided before or during an ongoing operation, but within 90 days after the end of the operation. The notice should include further information and an explanation of the remedy mechanisms (e.g., the legal framework and possible points of contact). Moreover, green space and potentially gray space targets should be notified *ex post*.

Oversight must be part of the legal framework.

4.2.3. Create transparency and auditability

Transparency about active cyber defense operations will enhance policies and controls. In the cybersecurity realm, there is a lot of expertise outside the government, in the private sector but also in academia and civil society. Disseminating, if necessary sanitized, details of active cyber defense operations beyond executive, judicial and legislative authorities allows external experts to independently assess measures and suggest improvements. Such details can be operational in nature or discuss the application of existing safeguards, a somewhat high-level version of the after-action

report. However, agencies should make an effort to provide transparency reports that are useful for external experts and published periodically, for example, once a year. Healey and Jervis remarked that

> *"[t]he national security community must declassify and break down compartments to combat cognitive bias. The current situation—yelping about the adversary's punches but classifying one's own—is not tenable, leading to a biased view of cyber conflict that is poisonous in an open democracy."*[95]

A key aspect of transparency as well as oversight is the auditability, performed by independent auditors, of the operations. Thus, audibility is only a prerequisite for transparency, not its substitute. Event logging and written statements from operators should form the basis for follow-on actions, such as after-action reports, and be accessible to the parliamentary oversight body. The technical side of the audit, therefore, should be as unchangeable as possible through a secured audit trail.

Due to the time aspect, *ex post* and *ex ante* auditing may be performed.

Transparency and auditability must be part of the legal framework.

4.3. Set up guidelines for tools and services

Although a vast range of measures falls under the active cyber defense operation definition used in this paper, some of them do not require additional tools. For example, mandating an ISP to reroute certain traffic would not require establishing additional capabilities. Some measures, however, especially the more intrusive ones, may require procurement. That could include software and services such as those procured by the German BKA in the disabling of Emotet malware or, in extreme cases, even exploits and intrusive tools.

Tools and services must be acquired only from third parties, ideally as open source, that are transparently vetted and which do not conduct any business with governments or other entities that have been reported to conduct unlawful activities and violate human rights with their tools and services.[96] In this respect, a due diligence audit must be conducted annually to avoid indirectly financing human rights abuses. A basis for this audit could be annual human rights reports from official and independent human rights bodies. For some countries, an alternative might be to develop such tools in-house[97] or share tools and exploits with allies. Whenever unknown vulnerabilities or exploits are required for an active cyber defense operation, a vul-

95 Jason Healey and Robert Jervis (2020), The Escalation Inversion And Other Oddities Of Situational Cyber Stability, Texas National Security Review
96 Aleksandra Sowa and Jan Mönikes (2012), Programmier- und Exportverbote für Software?, Frankfurter Hefte
97 Sven Herpig (2018), A Framework for Government Hacking in Criminal Investigations, Stiftung Neue Verantwortung

nerability assessment and management[98] should be established beforehand as a safeguard to manage the risks of reverse engineering, exploitation of leaks and/or use against the infrastructure by the threat actors[99]. Additionally, any intrusive tool must be thoroughly tested before deployment.[100]

4.4. Apply international law

Although it is, of course, a truism that international law must be obeyed within and outside blue space, the manner in which international law is applied to cyberspace is still evolving[101], including the use of countermeasures, the respective customary international law and potential *lex specialis*.[102] The international law framework for countermeasures seems appropriate for active cyber defense operations in red space, although they may not always qualify as countermeasures or reprisals under international law.[103] The framework includes aspects such as the measures should be proportional, have the right timing, be temporary, be non-retributive and require a certain level of attribution.[104] All these aspects must be considered for active cyber defense operations, as outlined in the criteria.

Governments leading active cyber defense operations must also be aware that they are contributing to the development of binding customary law. Merle Maigre cautioned that

> *"[w]hat like-minded states actually say on their understanding of international law matters a great deal. Over time, a critical mass of complementary state views on a particular cyber legal issue will accumulate—and that interpretation becomes a binding customary law, cementing the norms in place. Statements that are clearly expressing countries' interpretation on how the law is applying, help to clarify the legal framework where all of our nations operate."*[105]

That holds true for active cyber defense as well.

98 Sven Herpig (2018), Governmental Vulnerability Assessment and Management, Stiftung Neue Verantwortung

99 An interesting example of an unknown vulnerability that may have been exploited in active cyber defense operations to temporarily deny service to malicious cyber operation or campaign infrastructure and may not constitute a huge risk when not immediately disclosed to the vendor is Hotcobalt; see Gal Kristal (2021), Hotcobalt – New Cobalt Strike DoS Vulnerability That Lets You Halt Operations, SentinelOne and compare with Florian Roth (2021), Twitter; for a tool leveraging a different method for the same software, see Mario Henkel (2021), CobaltSpam, GitHub

100 Perri Adams, Dave Aitel, George Perkovich and JD Work (2021), Responsible Cyber Offense, LAWFARE

101 For example, Antonio Coco and Talita Dias (2020), The Oxford Process on International Law Protections in Cyberspace, University of Oxford

102 For example, Jack Goldsmith and Alex Loomis (2021), "Defend Forward" and Sovereignty, Hoover Institution and Ashley Deeks (2020), Defend Forward and Cyber Countermeasures, Hoover Institution and Henning Lahmann (2020), Unilateral Remedies to Cyber Operations, lehmann media and Michael Schmitt (2021), Three International Law Rules for Responding Effectively to Hostile Cyber Operations, JUST SECURITY

103 For example, Michael Schmitt (2021), Three International Law Rules for Responding Effectively to Hostile Cyber Operations, JUST SECURITY and Ashley Deeks (2020), Defend Forward and Cyber Countermeasures, Hoover Institution

104 For example, Janine Schmoldt (2020), Hacking Back aus völkerrechtlicher Perspektive, Defensive Con 2020 and Ashley Deeks (2020), Defend Forward and Cyber Countermeasures, Hoover Institution

105 Merle Maigre (2018), Diplomacy and Defense in Cyber Space, LAWFARE

4.5. Consider public interest

Not every malicious cyber operation or campaign against a company creates a demand for an active cyber defense operation. With limited response capabilities available to the government and limited scalability of active cyber defense operations[106], reacting to every breach would not be feasible.

Therefore, active cyber defense operations should be conducted only if there is a clear public interest in doing so, such as threats to public safety and security (thresholds may include interference with critical infrastructure or a local declaration of state of emergency), meddling with democratic processes or significant economic harm. If there is no clear public interest, then other, more passive, measures could be taken to neutralize or mitigate an ongoing cyber operation or campaign.

Enshrining this public interest threshold into an active cyber defense policy would constitute a red line that leaves enough strategic ambiguity for the implementing government to maneuver. In terms of transparency and confidence-building, it may be prudent to communicate this stance internationally, for example, in a cyber defense strategy or foreign policy statement.

4.6. Adapt confidence-building measures

If states consider active cyber defense operations in green, gray and/or red space, the states should set up international confidence-building measures or adapt existing ones for this purpose.[107]

An active cyber defense operation in green space should always be communicated to and approved by the respective government, *ex ante* if possible, as such an operation certainly "risks friction with allies"[108]. However, there may be an occasional need for *ex post* notification, for example, to avoid third-party collection. Especially for these cases, it is crucial to avoid misunderstandings, disgruntled allies and unnecessary escalation. The general intention to conduct active cyber defense operations in allied spaces should be discussed during bi- or multilateral talks and cyber dialogues. Which specific measure is chosen, for example, intelligence exchange, depends on the parties. This approach may also be applicable to gray space and even red space. The minimum viable communication should make sure that the targets of the active cyber defense operation understand *ex post* that it was a response and, therefore, had defensive intent. Needless to say, the chosen approach has to be in line with international law.

106 Matthias Schulze (2019), 05 /invite Sven Herpig – aktive Cyber-Abwehr, Hackback und Deutschlands Cyber-Sicherheitsarchitektur
107 For example, OSCE Secretariat (2021), Cyber/ICT Security, Organization for Security and Co-operation in Europe
108 Max Smeets (2019), Cyber Command's Strategy Risks Friction With Allies, LAWFARE

Apart from simple communication measures, states could, for example, build confidence through creating and adopting a common framework for active cyber defense operations, such as the one suggested in this paper. Another confidence-building measure could be joint cybersecurity policy exercises[109] with active cyber defense operations and stakeholders representing various spaces.

Whether confidence-building measures change the decision-making on the other end is up to the affected parties, but the measures should at least lessen the risk of escalation, although no opportunity to exercise due diligence first was given.

109 Rebecca Beigel and Julia Schuetze (2021), Cybersecurity Exercises for Policy Work, Stiftung Neue Verantwortung

5. Applications

5.1. Removal of the Hafnium web shells

5.1.1. Background

On March 2, 2021, Microsoft disclosed that a "state-sponsored threat actor" (Hafnium) operating from China had "engaged in a number of attacks using previously unknown exploits targeting on-premises Exchange Server software."[110] This and other malicious campaigns were able to intrude into and install web shells on the servers via ProxyLogon vulnerabilities.[111] At the same time, Microsoft released updates permitting the patching of these exploits. The update was followed up with a Joint Advisory issued by the FBI and the Cybersecurity and Infrastructure Security Agency (CISA) on March 10[112]. In addition, the White House urged Microsoft to provide "a second patch—a "one click" tool that was easier to deploy."[113]

Despite the availability of the patches and the advisory, "hundreds of vulnerable computers in the United States" were not patched, and the respective companies did not remove the web shells until the end of March.[114] The FBI requested a search-and-seizure warrant that would enable the agency to remotely remove the web shells, because the agency believed "that the owners of the still-compromised web servers did not have the technical ability to remove them on their own and that the shells posed a significant risk to the victim"[115] and more generally, "threaten[ed] the national security and public safety of the American people and our international partners."[116] The warrant was approved on April 9, 2021.[117]

Between April 9 and April 13, the FBI employed remote access methods to search and access previously identified file paths on servers in the United States[118] based on known, detected and commonly used passwords by the operators of the malicious cyber campaign. In the process, the agency created copies of the web shells for evidence and then "executed a command to uninstall the web shell from the

110 Tom Burt (2021), New nation-state cyberattacks, Microsoft
111 Tara Seals (2021), Microsoft Exchange Servers Face APT Attack Tsunami, Threatpost and Catalin Cimpanu (2021), FBI operation removed web shells from hacked Exchange servers across the US, The Record
112 Cybersecurity & Infrastructure Security Agency (2021), FBI-CISA Joint Advisory on Compromise of Microsoft Exchange Server, Cybersecurity & Infrastructure Security Agency
113 John Hudson and Ellen Nakashima (2021), U.S., allies accuse China of hacking Microsoft and condoning other cyberattacks, The Washington Post
114 U.S. Department of Justice (2021), Justice Department Announces Court-Authorized Effort to Disrupt Exploitation of Microsoft Exchange Server Vulnerabilities, U.S. Department of Justice
115 Lawrence Abrams (2021), FBI nuked web shells from hacked Exchange Servers without telling owners, Bleeping Computer
116 U.S. Department of Justice (2021), Justice Department Announces Court-Authorized Effort to Disrupt Exploitation of Microsoft Exchange Server Vulnerabilities, U.S. Department of Justice
117 U.S. District Court, South District of Texas, Houston Division (2021), CASE NO. 4:21mj755, U.S. Department of Justice
118 U.S. Department of Justice (2021), Justice Department Announces Court-Authorized Effort to Disrupt Exploitation of Microsoft Exchange Server Vulnerabilities, U.S. Department of Justice

compromised server."[119] The targets of this operation were informed only *ex post*.[120] No patches were rolled out during the operation[121], leaving the target systems free of the web shells but vulnerable to re-exploitation.

5.1.2. Criteria

Goal: The goal of the operation was to (temporarily) *mitigate* the impact of the malicious cyber campaign and give the affected organizations more time to patch as the systems had been compromised and remained vulnerable to re-exploitation after the active cyber defense operation.

Success: The web shells were removed, and the operators were informed about the urgency to patch. There are no public reports of the removal of the web shells causing disruptions or any other problems with the affected IT systems. The operation did not directly protect against re-exploitation of the vulnerability by the same threat actor or any other actor. Moreover, although the People's Republic of China was jointly called out for irresponsible state behavior in the aftermath by several countries[122], there is no evidence that the U.S. government needed the web shell removal for technical attribution. Therefore, the operation led to a *tactical success*.

Type: The operation was technically *reversible*, as the affected targets could theoretically reinfect their IT systems with the web shells[123], and *intrusive,* as it compromised the confidentiality and integrity of the target systems.

Space: The operation took place in *blue space,* as it was limited to IT systems located in the United States.[124]

Target: There was *no explicit mention of critical infrastructure*. The vulnerable software was on-premise Microsoft Exchange.

Government lead agency: The operation was led by the FBI, a *law enforcement* agency.

119 Lawrence Abrams (2021), FBI nuked web shells from hacked Exchange Servers without telling owners, Bleeping Computer
120 April Falcon Doss (2021), We're From the Government, We're Here to Help: The FBI and the Microsoft Exchange Hack, JUST SECURITY
121 U.S. Department of Justice (2021), Justice Department Announces Court-Authorized Effort to Disrupt Exploitation of Microsoft Exchange Server Vulnerabilities, U.S. Department of Justice
122 The White House (2021), The United States, Joined by Allies and Partners, Attributes Malicious Cyber Activity and Irresponsible State Behavior to the People's Republic of China, The White House
123 It is debatable, however, whether the FBI could have reversed the IT system state by reinstalling the web shells on the target systems if it came to that.
124 U.S. Department of Justice (2021), Justice Department Announces Court-Authorized Effort to Disrupt Exploitation of Microsoft Exchange Server Vulnerabilities, U.S. Department of Justice

Cooperativeness: The cooperativeness between the government lead agency and the targets of the active cyber defense operation is *unknown*. Although the targets of the active cyber defense operation were informed only after the fact, an indicator that rules out consent, they possibly might have been cooperative or even grateful[125], if they had been informed *ex ante*. The reason for the lack of *ex ante* notification might have been the urgency to remove the web shells before they could be exploited.

Attribution: Attribution was *not necessary* for the deployed measures, as they did not target the origin of the malicious cyber campaign but its victims. However, attribution was performed[126], which may have played a role in the decision about which space to conduct the active cyber defense operation in.

Time: The operation took place after the systems had been compromised but before they could be exploited further by the perpetrator or third parties. Therefore, the operation was carried out *during operations of the same campaign*, or at least *in-between sequential campaigns*.

De- and escalation: Considering the bluntness and targeting of the malicious cyber campaign and the restraint shown by the government agency through limiting its actions to removing the web shell in blue space, the operation was likely *de-escalating*, or rather *non-escalating*. In the way that proportionality is applied to the concept, *vis-à-vis* an adversary, it does not play a role in this operation.

Automation: The level of automation is unclear. In light of the hundreds of targets and the time frame of four days, even considering a preparation period, the operation was likely at least *semi-automated*, in the sense of tooling to enable scaled deployment to selected targets.

Frequency: It was a *one-off* operation. However, if the vulnerability is left unpatched, it cannot be ruled out that the agency may repeat the operation *periodically*.

Costs: Although the FBI contracted external expertise to test their tooling[127], it is *unlikely that the operation was costly*. For example, it did not require the procurement of exploits or expensive mockup target infrastructure.

Collateral consequences: Removing the web shells returned the IT systems to their pre-infection state by leveraging the access given to the government lead agency

125 April Falcon Doss (2021), We're From the Government, We're Here to Help: The FBI and the Microsoft Exchange Hack, JUST SECURITY
126 April Falcon Doss (2021), We're From the Government, We're Here to Help: The FBI and the Microsoft Exchange Hack, JUST SECURITY
127 April Falcon Doss (2021), We're From the Government, We're Here to Help: The FBI and the Microsoft Exchange Hack, JUST SECURITY

through the previously installed web shells. Additionally, an independent technical expert was engaged to assess the operation, possibly helping to rule out known unknowns and better track interdependencies.[128] Therefore, it is likely that *collateral consequences or known unknowns leading to collateral consequences were at least partially ruled out*.

Criterion	Assessment
Goal	Mitigation
Success	Tactical
Type	(Reversible) Intrusive
Space	Blue space
Target	Non-critical infrastructure (Critical infrastructure)
Government lead agency	Law enforcement
Cooperativeness	Unknown
Attribution	Not necessary
Time	During operations of the same campaign In-between sequential campaigns
De- and escalation	(Potential de-escalation) No change in the escalation cycle
Automation	Semi-automated
Frequency	One-off (Periodic)
Cost	Low
Collateral consequences	Not expected

Table 3. Active cyber defense criteria for removal of the Hafnium web shells

128 April Falcon Doss (2021), We're From the Government, We're Here to Help: The FBI and the Microsoft Exchange Hack, JUST SECURITY

5.1.3. Safeguards

Scope: The scope was clearly defined, targeting all national victims of the web shells that had not removed them and was kept within the government's jurisdiction.

National legal framework: The lack of discussion about the legality of the operation, even by critics, as well as a reform that addressed existing issues several years before, suggests that the legal framework in place was sufficient. However, no dedicated legal framework for active cyber defense operations exists, and it is unclear whether a legal analysis has been undertaken within the U.S. government.

Impact assessment: Although it is not clear whether the FBI made an impact assessment, the agency contracted an independent technical expert to evaluate the planned operation[129]. Therefore, this requirement seems to have been fulfilled, at least to a certain degree.

Oversight: The warrant, U.S. Department of Justice approval and *ex post* notification of the targets satisfy the oversight requirement, although an *ex ante* notification would have been more appropriate. It would have enabled the targets of the active cyber defense operation to either take precautions or remove the web shells before the government intervened.

Transparency and auditability: A sufficient amount of information about the operation was shared with the public after the operation. What information was shared with other relevant government stakeholders internally is not public knowledge. The same applies to the existence of a technical audit trail.

Procurement: Based on public information, no procurement of tools and services, apart from the independent evaluation of the operation, was required.

International law: The operation took place in blue space; thus, international law, in the sense of between nations, was not applicable.

Public interest: Taking into consideration the high number of targets affected by the Hafnium web shells (hundreds of IT systems), the subsequent implications if the targets were further exploited by threat actors and the debate in media outlets, the operation was likely in the public interest.

Confidence-building measures: The operation took place in blue space; thus, no international confidence-building measures were required for this operation.

129 April Falcon Doss (2021), We're From the Government, We're Here to Help: The FBI and the Microsoft Exchange Hack, JUST SECURITY

Safeguard	Assessment
Define and limit the scope	Clearly defined
Establish a national legal framework	Non-specific legal framework exists
Require impact assessments	Unclear; at least an independent technical expert was consulted
Implement oversight	*Ex ante* warrant and *ex post* notification of targets were required
Create transparency and auditability	Public was informed with a sufficient amount of information; auditability is unknown
Set up guidelines for tools and services	Not required
Apply international law	Not applicable (in the sense of between nations)
Consider public interest	Appears to have been in the public interest
Adapt confidence-building measures	Not required (in the sense of international measures)

Table 4.
Active cyber defense safeguards for removal of the Hafnium web shells

5.1.4. Assessment

The removal of the Hafnium web shells checks many of the right boxes. The law enforcement operation took place with a clear scope in blue space and with previous judicial authorization to mitigate further damage stemming from an ongoing malicious cyber campaign—and, therefore, was likely in the public interest. Although the FBI deployed (in an at least semi-autonomous way) intrusive measures that may have also affected critical infrastructure, the agency consulted with an independent technical expert before implementing the operation. From the risk and risk-mitigation point of view, the only complaint is the *ex post* notifications, which denied the targets, especially potential critical infrastructure, an opt-out or other precautions. How effective the operation was is more difficult to determine, but it was likely a tactical success. Although the operation removed only the web shells and did not patch the vulnerability (which would technically have been possible), it left the companies vulnerable to re-exploitation. However, the operation increased the threshold for that to happen. At the same time, the tools provided by the vendor were circulated by the government, and the targets of the web shell removal were informed; thus, they were aware and could patch their systems and infrastructure themselves.

Weighing the risks, risk mitigation and effectiveness of the operation based on public information, it seems that the removal of the Hafnium web shell was a responsible active cyber defense operations.

5.2. Disabling of Emotet malware

5.2.1. Background

On January 27, 2021, Europol announced that the "world's most dangerous malware Emotet" had been disrupted. The announcement was preceded by a collaborative multinational effort, also known as Operation Ladybird, that included the participation of authorities in the Netherlands, Germany, the United States, the United Kingdom, France, Lithuania, Canada and Ukraine under the coordination of the European Union Agency for Law Enforcement (Europol) and the European Union Agency for Criminal Justice Cooperation (Eurojust).[130] Operation Ladybird allowed government agencies to "gain control of the infrastructure and [take] it down from the inside"[131] and place the malicious software in quarantine within infected machines[132]. Then, the authorities also adjusted "communication parameters of the software [...] in a way that the victim systems no longer communicate with the infrastructure of the offenders but with an infrastructure created for the seizure of evidence."[133] The Dutch police implied that during the process they had "discovered and disrupted infrastructure backups" which would make resurrection of the Emotet botnet unlikely.[134] The Dutch policy also offers a form for checking whether an organization's email is part of the seized data, and it is likely that one or more of their IT systems was infected with Emotet malware.[135]

Operation Ladybird combined various elements within and outside cyberspace. The active cyber defense operation, which led to disabling of the malware, was led largely by the German BKA in cooperation with the German Central Office for Combating Internet and Computer Crime (ZIT), which secured the needed seizure warrant from a local court (Amtsgericht) in Gießen.[136]

However, whether this operation was lawful has been strongly disputed[137], mainly because the government claimed that the operation was to be only an evidence-collection operation for criminal prosecution, which would have been covered by the existing legal framework for the government lead agency. However, the federal gov-

130 European Union Agency for Law Enforcement Cooperation (2021), World's Most Dangerous Malware Emotet Disrupted Through Global Action, Europol
131 Andy Greenberg (2021), Cops Disrupt Emotet, the Internet's 'Most Dangerous Malware', WIRED
132 Sergiu Gatlan (2021), Emotet malware nukes itself today from all infected computers worldwide, Bleeping Computer
133 Sergiu Gatlan (2021), Emotet malware nukes itself today from all infected computers worldwide, Bleeping Computer
134 Andy Greenberg (2021), Cops Disrupt Emotet, the Internet's 'Most Dangerous Malware', WIRED
135 Dutch Police (2021), Controleer of uw e-mail en wachtwoord gestolen zijn door de Emotet malware, Politie
136 Sven Herpig and Dennis-Kenji Kipker (2021), German Emotet takedown in the legal gray zone, Stiftung Neue Verantwortung and Sergiu Gatlan (2021), Emotet malware nukes itself today from all infected computers worldwide, Bleeping Computer and Anna Biselli (2021), Darf das BKA Schadsoftware auf infizierten Rechnern manipulieren?, Netzpolitik.org
137 Dennis-Kenji Kipker and Michael Walkusz (2021), Das BKA zerschlägt die Infrastruktur von Emotet: Mit welcher Rechtsgrundlage?, beck-community and Sven Herpig and Dennis-Kenji Kipker (2021), German Emotet takedown in the legal gray zone, Stiftung Neue Verantwortung and Andre Meister (2021), BKA nutzt Emotet-Takedown als Türöffner für mehr Befugnisse und neue Gesetze

ernment lead agency also disabled the malware, which can be seen as an emergency response to a cybercrime, a power not granted to that agency.

With technical support from a German IT security company, the BKA rolled out a software update (module) to more than 53,000 IT systems worldwide that placed the Emotet malware in quarantine, disconnected the target system from the Emotet botnet, sent back data (IP address, computer name and running processes) for evidence collection and effectively disabled the malware.[138] No patches were rolled out during the operation, leaving the target systems free of Emotet malware but vulnerable to re-exploitation.

Targets of the active cyber defense operation were informed only *ex post* through their ISPs, a process led by Germany's national cyber security agency (BSI) with information from the BKA.[139]

5.2.2. Criteria

Goal: The goal of the operation was to *mitigate* the impact of the malicious cyber campaign, as the systems had been compromised and remained vulnerable to re-exploitation after the active cyber defense operation. Although it remains unclear whether the disabling of the malware was covered by the existing legal framework, the patching of the systems would have definitely been outside it.

Success: The malware was disabled, and the operators of the respective IT systems were informed. Therefore, although the IT systems were still vulnerable, the victims could not be harmed directly if the Emotet operators had a command-and-control backup infrastructure. Although it is debatable whether Operation Ladybird as such was a tactical or strategic success, with other threat actors possibly sweeping in and compromising the still-vulnerable systems, the disabling of malware part was *a tactical success*.

Type: The operation was *intrusive*, as it compromised the integrity and confidentiality of the target systems. The operation was technically *reversible*, as the affected targets could theoretically reinfect their IT systems with Emotet malware.[140] Additionally, the module function which allowed sending back data was deactivated later.

138 U.S. Department of Justice (2021), Emotet Botnet Disrupted in International Cyber Operation, U.S. Department of Justice and Sergiu Gatlan (2021), Emotet malware nukes itself today from all infected computers worldwide, Bleeping Computer
Andre Meister (2021), BKA nutzt Emotet-Takedown als Türöffner für mehr Befugnisse und neue Gesetze

139 Andre Meister (2021), BKA nutzt Emotet-Takedown als Türöffner für mehr Befugnisse und neue Gesetze

140 The BKA, however, may not have been able to reverse it, as the agency legally would not have been allowed to re-deploy the original Emotet malware module to the target systems.

Space: As the module was deployed to more than 53,000 IT systems worldwide, it can be assumed that the operation took place in blue space and green space. Whether the operation was limited to non-red-space systems, or what the scope was in general (bearing in mind that around that time there may have been more than 1.6 million Emotet-infected IT systems[141]) is impossible to say with the publicly available information.

Target: There was *no explicit mention of critical infrastructure*. The module was rolled out to Microsoft Windows machines.

Government lead agency: The operation was led by the BKA, a *law enforcement* agency.

Cooperativeness: The cooperativeness is *unknown* and *noncooperative*. Although the targets of the active cyber defense operation were informed only after the fact, an indicator that rules out consent, they might have been cooperative if they had been informed *ex ante*. That is more likely to be true for those in blue space than for those in green or even gray space.

Attribution: Technical attribution was *not necessary* for the deployed measures, as they did not target the origin of the malicious cyber campaign but its victims. The parameters to disconnect the targets from the botnet were provided by other activities of Operation Ladybird.

Time: The operation took place after the systems had been compromised and then infected with malware but before the malware was activated by the threat actor or third parties to have severe effects on the target. Therefore, the operation took place either *during operations of the same campaign* or *in-between sequential campaigns*.

De- and escalation: Considering the damage that Emotet has caused over the years, and that it is a cybercrime group and not a government threat actor, it is *unlikely that there was a change in the escalation cycle* when looking at Operation Ladybird overall. Limiting the scope to the active cyber defense operation led by the BKA (which, of course, would have not been possible without the access gained through other parts of Operation Ladybird) may even have been *de-escalatory*. In the way that proportionality is applied to the concept, *vis-à-vis* an adversary, it did not play a role in this operation.

141 Andre Meister (2021), BKA nutzt Emotet-Takedown als Türöffner für mehr Befugnisse und neue Gesetze

Automation: The level of automation is unclear. Considering the tens of thousands of targets and the time frame, even considering a preparation period, the operation was likely at least *semi-automated* (deployment of software).

Frequency: It was a *one-off* operation. Although previous attempts to disrupt botnets have not been sustainable, this one seems to be sustainable (thus far[142]), also because backup structures were targeted in the larger operation.

Costs: Although the BKA contracted external expertise to develop and deploy the tooling, it is *unlikely that the operation was costly*. For example, it did not require the procurement of exploits or expensive mockup target infrastructure.

Collateral consequences: Disabling the malware partially returned the IT systems to their pre-infection state by leveraging the access given to the government lead agency through the command-and-control infrastructure by which the malware was controlled. Additionally, the government lead agency cooperated with a company with in-depth knowledge of the threat actor and its technical infrastructure, possibly helping to rule out known unknowns and better track interdependencies. Therefore, it is likely that *collateral consequences or known unknowns leading to collateral consequences were at least partially ruled out*.

Criterion	Assessment
Goal	Mitigation
Success	Tactical
Type	(Reversible) Intrusive
Space	Blue space Green space (Gray space)
Target	Non-critical infrastructure (Critical infrastructure)
Government lead agency	Law enforcement

142 Andy Greenberg (2021), Cops Disrupt Emotet, the Internet's 'Most Dangerous Malware', WIRED

Criterion	Assessment
Cooperativeness	Unknown Noncooperative
Attribution	Not necessary
Time	During operations of the same campaign In-between sequential campaigns
De- and escalation	No change in the escalation cycle
Automation	Semi-automated
Frequency	One-off
Cost	Low
Collateral consequences	Not expected

Table 5. Active cyber defense criteria for disabling of Emotet malware

5.2.3. Safeguards

Scope: The scope of the operation remains unclear, as it apparently targeted fewer systems than were compromised in Germany. At the same time, the operation targeted systems outside Germany, where the BKA has no jurisdiction. Without additional information, which has not been provided by the government, the scope appears arbitrary. To better limit the scope, the BKA should have at least focused on blue space and handed over tools and access to the agency's green space allies to conduct similar operations in their respective countries.

National legal framework: The discussion about the legality of the operation, which led to a debate in the German parliament[143], shows that there is much work to be done concerning the legal framework. If the legal framework is not changed, operations like this one can be implemented only in terms of criminal prosecutions. That a debate about the legality of such operations would take place must have been clear, as it has been discussed for many years.[144]

Impact Assessment: It is unclear whether an *ex ante* impact assessment was conducted.

143 Andre Meister (2021), BKA nutzt Emotet-Takedown als Türöffner für mehr Befugnisse und neue Gesetze, Netzpolitik.org
144 Sven Herpig et al. (2020), Aktive Cyberabwehr/ Hackback in Deutschland, Stiftung Neue Verantwortung

Oversight: The BKA and ZIT obtained a warrant and informed the targets of the operation *ex post* through the national cybersecurity agency and the respective ISPs. Whether an *ex ante* notification would have been possible or the window of opportunity was closing is impossible to assess with the information available.

Transparency and auditability: The information shared directly with the public could be improved, especially with regards to technical details and which stakeholders were included in the decision-making process. What information was shared with other relevant government stakeholders internally is not public knowledge. The same applies to the existence of a technical audit trail.

Procurement: Tools and services were provided by a German company with a public track record that is, based on publicly available information, aligned with the parameters described in this safeguard.

International law: The operation was part of a joint operation. Thus, the scope is unclear and similar whether all the targets were based in countries that were part of the broader operation. Therefore, it is unclear whether the operation was aligned with international law.

Public interest: Taking into consideration the general threat of ransomware for the country, and the damage caused by Emotet estimated to total at least 14.5 million Euros in Germany[145] and hundreds of millions worldwide[146], it is safe to assume that the operation was in the public interest.

Confidence-building measures: This operation was part of a joint operation. Thus, the scope is unclear and similar whether all the targets were based in countries that were part of the broader operation. Therefore, it is also unclear if targets were based in countries that were not part of the joint operation and whether they were informed *ex ante* or during the operation as a confidence-building measure.

145 Bundeskriminalamt (2021), Infrastruktur der Emotet-Schadsoftware zerschlagen, Bundeskriminalamt
146 U.S. Department of Justice (2021), Emotet Botnet Disrupted in International Cyber Operation, The United States Department of Justice

Safeguard	Assessment
Define and limit the scope	Unclear
Establish a national legal framework	Applicability of existing non-specific legal framework questionable
Require impact assessments	Unclear
Implement oversight	*Ex ante* warrant and *ex post* notification of targets were required
Create transparency and auditability	Public was provided with a minimum of information; auditability unknown
Set up guidelines for tools and services	Likely implemented
Apply international law	Unclear (Not applicable (in the sense of between nations))
Consider public interest	Appears to have been in the public interest
Adapt confidence-building measures	Unclear if required (in the sense of international measures) but if required beyond partner states, unclear if implemented

Table 6.
Active cyber defense safeguards for disabling of Emotet malware

5.2.4. Assessment

Unfortunately, key information that would help to better assess the risk and risk mitigation (e.g., whether all countries in which systems were targeted approved this operation *ex ante*) was not disclosed by the government publicly.

A semi-automated, intrusive operation possibly including critical infrastructure without *ex ante* notification to the targets, and therefore, without an opt-out option, is risky, especially because the operation was not limited to blue space, and whether approval for green and gray spaces was given is unknown. These aspects are further exacerbated by the fact that although a warrant was procured, legal and policy scholars doubt the legality of this operation. *Ex post* notification informed the targets that they are still vulnerable and, therefore, can take appropriate actions. It is unclear whether and when the corresponding states the targets reside in were informed.

How effective the operation was is more difficult to determine. It disabled the malware but did not patch the underlying vulnerabilities; therefore, re-exploitation was still possible, especially by other threat actors. Compared to the *removal of the Haf-*

nium web shells, however, remote patching would not have been feasible because the Emotet infections could have occurred through various attack vectors and vulnerabilities, depending on the victim. The disabling of the Emotet malware increased the threshold for re-exploitation and neutralized a direct risk of the malware being activated to, for example, roll out ransomware. It may however be difficult to argue that the risks are justified by the operation's success, as it was merely tactical.

The political aftermath showed that the operation benefited from a kind of survivorship bias and the success of the broader Operation Ladybird. However, this should not distract from the risk based on the assessment of the criteria, for example in terms of intrusiveness without *ex ante* information and possible geopolitical implications, as well as non-sufficient risk-mitigation safeguards. For the latter, in particular, the lack of a clear legal framework, even to the point of the questionable legality of the operation, and transparency are problematic.

Based on the publicly available information, the assessed criteria and safeguards lead to the conclusion that although the disabling of the Emotet malware was operationally successful, it is unlikely that the measure was a responsible active cyber defense operation.

6. Recommendations

Active cyber defense operations include a wide range of measures. Although some measures can be extremely risky, especially when carried out in green or gray spaces, they may offer additional options to what is on the table. Considering the criteria and safeguards, understanding whether it would be responsible to conduct a certain active cyber defense operation requires a nuanced case-by-case approach.

A clear goal should be that the operation aims to neutralize, mitigate and/or technically attribute a specific malicious cyber operation or campaign and is not implemented for retribution. The crucial preconditions that enable active cyber defense operations to be legitimate options are a clear national legal framework, appropriate oversight mechanisms and an *ex ante* impact assessment that includes at the bare minimum the criteria discussed in this paper. Although these aspects should be considered on a case-by-case basis, they should be enshrined in a clear and structured process that includes additional stakeholders from the operational and strategic levels as well as from industry, academia and civil society. Setting up such a process could be facilitated by conducting a series of cybersecurity policy exercises[147]. Additionally, active cyber defense operations should be implemented in line with the public interest.

Based on the discussed aspects, the most sensible approach would be to first focus active cyber defense operations on what can be done in blue space by law enforcement and national cybersecurity agencies. At the international level, this approach should be supported by ensuring that more states agree on the existence of a violation of sovereignty through cyber operations and campaigns as well as the applicability of due diligence; the further development of customary international law through *opinio juris* and state practice in this sector[148]. Although development in the area of sovereignty violation through cyber operations could enable countermeasures in red space, development of due diligence obligations would possibly extend the scope to green and gray spaces as well.

Last, whether like-minded states would want to agree on a common understanding of active cyber defense operations and advance the dialogue on collective active cyber defense should be explored further[149], especially in light of malicious cyber operations or campaigns such as the Hafnium-ProxyLogon exploitation and the Emotet botnet. This includes not only additional dialogue on the topic but also possibly a common framework and joint active cyber defense exercises.

147 Rebecca Beigel and Julia Schuetze (2021), Cybersecurity Exercises for Policy Work, Stiftung Neue Verantwortung
148 Michael Schmitt (2021), Germany's Positions on International Law in Cyberspace Part I, JUST SECURITY
149 Ashley Deeks (2020), Defend Forward and Cyber Countermeasures, Hoover Institution and Michael Schmitt (2021), Three International Law Rules for Responding Effectively to Hostile Cyber Operations, JUST SECURITY

In conclusion, active cyber defense operations employing carefully selected and weighed measures with strong safeguards may increase national cybersecurity. However, it is crucial to integrate those measures into the existing cybersecurity culture, architecture and strategy. That requires open and nuanced technical and policy debates on use cases, criteria and safeguards, so that *imposing costs* does not just become "the government/ SIGINT agency equivalent of *we take your security very seriously*."[150] Ultimately, active cyber defense operations will be only a small fraction of activities that increase overall cybersecurity; IT security and resilience are still the king and queen.

150 Patrick Gray (2021), Risky Biz Feature Podcast: An interview with Rob Joyce, Risky Business

Acknowledgment

This analysis has been supported by the Transatlantic Cyber Forum working group on active cyber defense through online collaboration and a joint virtual workshop.

The views and opinions expressed in this paper are those of the author and do not necessarily reflect the official policy or position of the working group members or that of their respective employer/s.

In alphabetical order, acknowledging essential contributions of:

1. Dave Aitel, Cordyceps Systems
2. Charles-Pierre Astolfi, Advisor to the French Minister for Digital Affairs
3. Manuel Atug, HiSolutions AG
4. Ralf Benzmüller, G DATA CyberDefense AG
5. Sneha Dawda, Royal United Services Institute (RUSI)
6. Lars Fischer, Bremerhaven School of Applied Science
7. Stefanie Frey, Deutor Cyber Security Solutions
8. Kenneth Geers, NATO Cooperative Cyber Defence Centre of Excellence (NATO CCD COE)
9. Max Heinemeyer, Darktrace
10. Wyatt Hoffman, Center for Security and Emerging Technology (CSET), Georgetown University
11. Louise Marie Hurel, London School of Economics and Political Science (LSE)
12. Carolin Kemper, German Research Institute for Public Administration (FÖV)
13. Karsten König, Independent Security Researcher
14. Lucie Krahulcova, Digital Rights Watch
15. Andreas Kuehn, Observer Research Foundation America
16. Joanna Kulesza, University of Lodz
17. Thomas Lawson, axa
18. Sönke Marahrens, European Centre of Excellence for Countering Hybrid Threats
19. Igor Mikolic-Torreira, Center for Security and Emerging Technology (CSET), Georgetown University
20. Daniel Moßbrucker, University of Hamburg
21. Lukasz Olejnik, Independent Researcher and Consultant
22. Jörg Pohle, Alexander von Humboldt Institute for Internet and Society (HIIG)
23. Johanna Polle, Institute for Peace Research and Security Policy at the University of Hamburg (IFSH)
24. Thomas Reinhold, Chair of Science and Technology for Peace and Security (PEASEC), Technical University of Darmstadt
25. Christine Runnegar, Internet Trust, Internet Society
26. Janine Schmoldt, University of Erfurt

27. Emma Schroeder, Atlantic Council
28. Julia Schuetze, Stiftung Neue Verantwortung
29. Matthias Schulze, Stiftung Wissenschaft und Politik
30. Ari Schwartz, Center for Cybersecurity Policy and Law
31. Aleksandra Sowa, German Informatics Society, LG PET
32. Timo Steffens, German Federal Office for Information Security (BSI)
33. Eric Wenger, Cisco Systems

Additional thanks to workshop moderator Christina Rupp and the Stiftung Neue Verantwortung for their support.

About the Stiftung Neue Verantwortung

The Stiftung Neue Verantwortung (SNV) is an independent think tank that develops concrete ideas as to how German politics can shape technological change in society, the economy and the state. In order to guarantee the independence of its work, the organisation adopted a concept of mixed funding sources that include foundations, public funds and businesses. Issues of digital infrastructure, the changing pattern of employment, IT security or internet surveillance now affect key areas of economic and social policy, domestic security or the protection of the fundamental rights of individuals. The experts of the SNV formulate analyses, develop policy proposals and organise conferences that address these issues and further subject areas.

About the Transatlantic Cyber Forum (TCF)

The Transatlantic Cyber Forum (TCF) was established by the Berlin based think tank Stiftung Neue Verantwortung (SNV) in January 2017. The Transatlantic Cyber Forum is a network of cyber security experts and practitioners from civil society, academia and private sector. It was made possible with the financial support from the Robert Bosch Stiftung and the William and Flora Hewlett Foundation.

About the Author

Dr. Sven Herpig is the director for international cyber security policy at Stiftung Neue Verantwortung. His focal areas include information security of machine learning, (geopolitical) responses to cyber operations, government hacking and vulnerability management, and Germany's cybersecurity policy. Before Sven joined the Stiftung Neue Verantwortung, he was employed by Germany's federal government for several years.

Contact the Author

Dr. Sven Herpig
Director for International Cybersecurity Policy
sherpig@stiftung-nv.de
Twitter: @z_edian
T: +49 (0) 30 81 45 03 78 91

Imprint

Stiftung Neue Verantwortung e.V.
Beisheim Center
Berliner Freiheit 2
10785 Berlin

T: +49 (0) 30 81 45 03 78 80
F: +49 (0) 30 81 45 03 78 97

www.stiftung-nv.de
info@stiftung-nv.de

Design:
Make Studio
www.make-studio.net

Layout:
Jan Klöthe