CyberPeace Institute

cyberpeace watch

Report of Expert Meeting on the development of a Harms Methodology

A standardized methodology to measure the harms and impacts of cyberattacks and incidents on people, society and the environment.

The positions expressed in this Report are those of the authors in the CyberPeace Institute and do not represent the views or positions of any of the experts that took part in consultations. The content in this Report will be the subject of consultations and as such may be updated or modified in the coming months.

The development of the Harms Methodology is an ongoing process, and thus ahead of its finalization in 2024, it is respectfully requested that the contents of this Report are not used or cited without the express permission of the authors. All content, including text, images, logos, and graphics, in this Report are the property of the of the CyberPeace Institute.

Report Structure

Executive Summary

The frequency, scope, sophistication, and severity of cyberattacks and cyber incidents have increased at an alarming pace in recent years, and will continue to do so, exposing vulnerable communities. Whether in peacetime or war it is important that in our technology-dependent world there is a recognition that cyberattacks do not just attack or harm technology, do not always have (easily) reversible effects, and can have impacts at national and international levels. Efforts to measure these impacts have focused on the direct impact to targeted systems or organizations, this affects the ability to understand and measure the extent of the actual harm caused to people, society and the environment. This impedes policy making, resilience efforts and a means to affirm the real harm of a cyberattack for victims, including in accountability processes.

A clarification on what constitutes harm in a comprehensive and measurable manner is thus required, coupled with data-driven and evidence-based metrics, tools and frameworks for understanding, tracking, and measuring this harm. Recognizing this, the CyberPeace Institute, research and a process to develop a harms methodology. The strategic objective is to determine the means to measure harm from cyberattacks and incidents in order to increase knowledge of the human costs, and influence policy, accountability and resilience efforts.

The Institute held a first Expert Meeting, a multistakeholder workshop, on 7th November 2023, to share its progress and to gather expert insights on a draft harms methodology to measure the harms and impacts of cyberattacks and incidents on people and society. The aim was to stress test the work carried out to date, and to gather insights for its evolution. This Report is a summary of the detailed observations and recommendations provided during this workshop, and includes how this work is guiding or reflected in the ongoing work to develop this methodology. This Report will also be the basis for continued consultation of experts and other stakeholders over the coming months.

The CyberPeace Institute has determined that a Theory of Violence - which implies an intention of harm - is a valid analytical tool for any kind of violence and is valid for analyzing cyberattacks and incidents.

The Institute defines harm as:

- A negative impact on the victim or victims' **physical, psychological, social, economic** well-being, their physical security, their economic security, or on the **environment**.

The Institute determined the following definition of Cyber Violence as:

- *The purposeful use, threat of use, negligent use or autonomous action*[1] *of digital*

and information technologies that directly, indirectly, temporarily or permanently causes either immediate or long-term harm, determined as negative impact on people's health[2], their physical security, their economic security or on the environment.

The CyberPeace Institute proposes a typology of harm caused by cyberattacks or cyber incidents, where the initial direct impact must be at a digital level or come from digital means. This can then have an additional or spillover impact (indirect impact) on people or the environment. Alternatively - and much more likely, - the impact on people or the environment is caused by or resulting from digital means or the digital impact on security, services, institutions, finances, or the economy.

Definitions, proposed in a Lexicon, and the typology of harm help to qualify the impact of cyberattacks and incidents and the associated risk factors. For mitigation, prevention and accountability, the next step is quantifying the impact.

Categorizing harms is an important part of the definition of cyber violence. The Institute found it prudent to simplify these to the following categories which encompass the holistic range of harms concisely, and in order to enable meaningful measurement: Physical, Psychological, Social, Deprivational, and Environmental. Robust categories of harm will be developed that are clearly distinct and not overly inclusive. The Institute will explicitly clarify categories and harmonize them with definitions. The Lexicon will explain what is included in each category. A draft lexicon has been developed and will be the subject of consultations over the coming months.

The Harms Methodology is being constructed according to the following principles:

1. The Methodology is adaptable to new developments and lessons learned; and developed alongside further research into the mechanisms and definitions of harms from the use of cyber.
2. Metrics are chosen with a consideration of the context of the attack or operation, (target, method, industry, etc.).
3. The Methodology will include specific guidance on data collection.
4. A holistic methodology faces inherent challenges if it aims to measure different kinds of harm under one standard, e.g. summarizing both physical and psychological harm with one figure.

Case studies were presented to experts outlining the extraction of indicators of harm. Learnings from the aggregation of metrics for an examination of case studies are being elaborated and will be shared as part of the explanation of the development of the Methodology. The Institute is following up on research further to questions elaborated by the experts, particularly: Do we start with the categories, and create indicators, or

do we start with indicators and create categories? Is this influenced by what data is available? The Institute will continue its research into the matter of providing a scoring or weighting within the categories of harm and the validity or not of comparing harms across categories. This will be modeled using case examples.

The Expert Meeting was an important first milestone for the CyberPeace Institute in publicly sharing its work to develop a standard data driven harms methodology and metrics to understand, track, and measure the harm from cyberattacks and incidents. The meeting allowed the CyberPeace Institute to confirm much of its research, to nuance some of its thinking benefitting from the feedback received, and to move forward and confirm next steps.

The complexity of the development of such a methodology and metrics is important to underline due to the broad range of considerations that need to be factored into this work. However, the important contribution that such a Methodology could make to understanding and measuring harm more comprehensively was underlined.

The publication of this Report of the Expert Meeting will enable the Institute to engage with experts who could not attend this meeting, and to broaden our outreach to a range of additional stakeholders for their insights. The Institute will leverage this Report to engage with States and civil society actors over the next months.

In parallel, the Institute will continue its research focusing particularly on an ontology of terms for data collection, and operationalising the definitions through continued work on indicators and metrics, including through assessments of further case studies and a range of types of cyberattacks and incidents. Case studies continue to enable the Institute to explore indicators and metrics, and to test data collection needs.

In this regard, the Institute is also working on a pilot project and AI modeling based on known features of the harm caused by a cyberattack or a cyber incident- together with other details such as claims by perpetrators or threat actors. This modeling entails leveraging AI as a diagnostic tool that then gives possibilities of type of attack, speed of spread, the “knock-on” human impact, origin, type of attack, intent, etc. The focus will be on instructing the tool to undertake the analysis and write the outcome in the format given by the definition of the Theory of Violence. The findings of this research will be consolidated into background documents for a further consultation meeting with experts.

The Institute aims to convene a second Expert Meeting in the second quarter of 2024 in order to present developments in this work and seek further insights and recommendations. Meanwhile, any feedback on this Report and ongoing work can be shared with the Report authors. The Institute welcomes engagement and collaboration on this work.

Introduction

The frequency, scope, sophistication, and severity of cyberattacks[3] and cyber incidents have increased at an alarming pace in recent years, and will continue to do so, exposing vulnerable communities. Whether in peacetime or war, - or the perceived gray zone[4] between the two - it is important that in our technology-dependent world there is a recognition that cyberattacks do not just attack or harm technology, do not always have (easily) reversible effects, and can have impacts at national and international levels. A clarification on what constitutes harm in a comprehensive and measurable manner is thus required.

In relation to the aim of cyberattacks or cyber incidents, terms used are effects, disrupt, degrade, destroy, deceive, deny, dysfunction, and exfiltrate. The use of cyber means to disinform, i.e. to further information operations is also important to underline. In relation to the exposure of communities to cyberattacks, many different terms are used - often interchangeably - to explain the resulting consequences, results, effects, impact, outcome, damage and harm to the victims of such attacks.

Efforts to measure these consequences have focused on the direct impact to targeted systems or organizations; from time to restore, financial costs and to some extent the number of breached records. This affects the ability to understand and measure the extent of the actual harm caused to people, society and the environment.

There is currently no standard methodology and a lack of metrics, tools and frameworks for understanding, tracking, and measuring this harm. A data-driven and evidence-based approach to measuring the harm from cyberattacks and cyber incidents is needed now more than ever.

Recognizing this, the CyberPeace Institute initiated, in 2022, research and a process to develop a harms methodology. The strategic objective is to determine the means to measure harm from cyberattacks and incidents in order to increase knowledge of the human impact, empower victims, and influence policy, accountability and resilience efforts.

The harms methodology is a key contribution to the Institute's CyberPeace Watch program. The CyberPeace Watch aims to provide a publicly accessible baseline of data to understand and share knowledge about cyberattacks, including threat analysis, harm, applicable laws and norms, and related paths for accountability. The platform's goal is to assess cyber peace based on evidence of the harm caused by cyberattacks and the actions taken by states and other relevant actors to strengthen responsible behavior in cyberspace. The Platform will launch in 2024.

Why is it important to measure harm?

A narrow definition of violence, to only physical violence, and a lack of knowledge of the harm of cyberattacks and incidents undermines a true evaluation of the scope and magnitude of such attacks. This then impedes policy making, resilience efforts and a means to affirm the real harm of a cyberattack or a cyber incident for victims, including in accountability processes.

The 2021 Report of the United Nations Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security, (UN GGE)[5], calls for States to further advance transparency and predictability including through voluntary sharing by States of e.g. *"national approaches to classifying incidents in terms of the scale and seriousness of the incident"*[6], and *"…. frameworks … for identifying, classifying and managing ICT incidents affecting critical infrastructure"*[7].

Eleven UN norms were first agreed upon by a UN GGE in 2015, with its reports endorsed by consensus at the UN General Assembly through resolution 70/237.[8] The normative framework for responsible state behavior in cyberspace aims to reduce risks to international peace and security, and to contribute to conflict prevention. These voluntary, non-binding norms outline both positive obligations and negative obligations with regards to how states should act in cyberspace, eight relate to actions that states want to encourage and three involve actions that countries should avoid.

The UN Open-ended Working Group on security of and in the use of Information and Communications Technologies (UN OEWG) focused on the operationalisation and implementation of the UN norms. Several of the norms refer to relevant terms for an analysis of harm (emphasis added):

- *"a. … States should cooperate in developing and applying measures to increase stability and security in the use of ICTs and to prevent ICT practices that are acknowledged to be harmful or that may pose threats to international peace and security;*
- *b. In case of ICT incidents, States should consider all relevant information, including the larger context of the event, the challenges of attribution in the ICT environment and the nature and extent of the consequences; …*
- *f. A State should not conduct or knowingly support ICT activity contrary to its obligations under international law that intentionally damages critical infrastructure or otherwise impairs the use and operation of critical infrastructure to provide services to the public; ..*
- *i. States should take reasonable steps to ensure the integrity of the supply chain … States should seek to prevent the proliferation of malicious ICT tools and techniques and the use of harmful hidden functions; …*
- *k. States should not conduct or knowingly support activity to harm the information systems of the authorized emergency response teams … of another State… …".*[9]

The United Nations Institute for Disarmament Research (UNIDIR) has provided in 2022 a framework document[10] which elaborates the "Foundational Cyber Capabilities" relevant for States to implement the 11 norms of responsible behavior. This includes:

- in relation to Norm B having a "Classification (public or non-public) of ICT incidents in terms of scale and impact"[11], and
- in relation to Norm F and G having a "Classification (public or non-public) of ICT incidents in terms of scale and seriousness"[12].

UNIDIR also published in 2022 a Taxonomy of Malicious ICT Incidents[13] which focuses on disruptive effects, considered as *"effects generated by the disruption of operations (such as message manipulation, denial of services, and data attacks)"*. It does not include exploitative effects, considered as *"the effects that result from incidents aimed at stealing information (such as exploitation of network infrastructure, or exploitation of data in transit)"*.[14] This Taxonomy leverages the 2016 work of Ioannis Agrafiotis, et al. on cyber harm[15].

A standard data driven harms methodology and metrics to understand, track, and measure the harm from cyberattacks and cyber incidents could support the above ambitions. The harms methodology will be provided to policy makers to contribute to evolving policy negotiations and to practitioners focused on building stronger accountability measures. The Institute will also use the methodology in its own work to track and measure the harms from cyberattacks and to call for responsible behavior in cyberspace.

Purpose and modalities of the Expert Meeting/Workshop

The Institute held a first Expert Meeting, a multistakeholder workshop, on 7th November 2023, to share its progress and to gather expert insights on a draft harms methodology to measure the harms and impacts of cyberattacks and incidents on people and society. The aim was to stress test the work carried out to date, and to gather insights and recommendations for its evolution. This included presenting an extrapolation of indicators of harm from two case studies.

The meeting was held under the Chatham House Rule, and in hybrid format, with thirty five participants contributing either in person in Geneva or online. See Annex 1 for a full list of participants who were drawn from International Organizations (IOs), Non Governmental Organizations (NGOs), legal, technical and academic communities. Experts were guided by a number of questions to discuss in both plenary and group working sessions with rapporteurs from the Institute collecting feedback. This feedback was summarized in several slides presented to all participants at the close of the meeting.

This Report is a summary of the detailed observations and recommendations provided during this workshop, and includes additional research suggested during the Meeting, and includes how this work is guiding or reflected in the ongoing work to develop this methodology.

For ease of reading, the Report sections are color coded as follows:

- Green sections: Content, research and conclusions of the CyberPeace Institute
- Yellow section: Feedback provided by participants at the Expert Meeting
- Red section: Terms and legal definitions

The CyberPeace Institute sincerely thanks the experts who participated in the meeting for generously providing their expertise, time and insights. The views expressed in this Report are those of the authors.

Definitions of violence and harm

Definitions are important to clarify and demarcate scope, methodology, and analysis, including for the data to collect and aggregate.

Defining a theory of violence in relation to cyberattacks

Increasingly an understanding of cyberattacks encompasses more than the means of an attack or its immediate impact. The manner of commission of an attack or incident, the scale and nature of the attack as well as the targeted victim all have a bearing on an understanding of its gravity.

The CyberPeace Institute proposes that a Theory of Violence - which implies an intention of harm - is a valid analytical tool for analyzing cyberattacks and incidents. The Institute determined the following definition of Cyber Violence[16] as:

"The purposeful use, threat of use, negligent use or autonomous action[17] of digital and information technologies that directly, indirectly, temporarily or permanently cause either immediate or long-term harm, determined as having a negative impact on people's health[18], their physical security, their economic security or the environment."

Feedback at the Expert Meeting and the work of several organizations were useful in determining the above definition. In particular, language from the Council of Europe Cybercrime Convention Committee[19], a statement of the United Nations Special Rapporteur on Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment[20], and the World Health Organization (WHO) definitions.

WHO states that *"Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity."* [21], and defines violence as *"the intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community, that either results in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment, or deprivation."* [22]

The Institute also leveraged the work of several academics[23] who had previously combined these two WHO definitions to build a "Theory of Violence" which provides a "harm-oriented" analytical tool

for any kind of violence. This theory defines violence as: *"The intentional use or threat of physical force against a person, group or community, that has a negative impact on the victims' physical, psychological or social wellbeing including deprivation."*[24] [25]

Impact here is taken to broadly describe the full range of effects of an incident, including impacts that do not cause harm, impacts that do cause harm, and harm itself. Impact may be understood as effects on any of the three sections of the Typology of Harm graphic in the following section, (e.g digital impact, impact to security, services and institutions, and to people, society and the environment.)

Definition and Typology of harms

The Institute defines harm as:

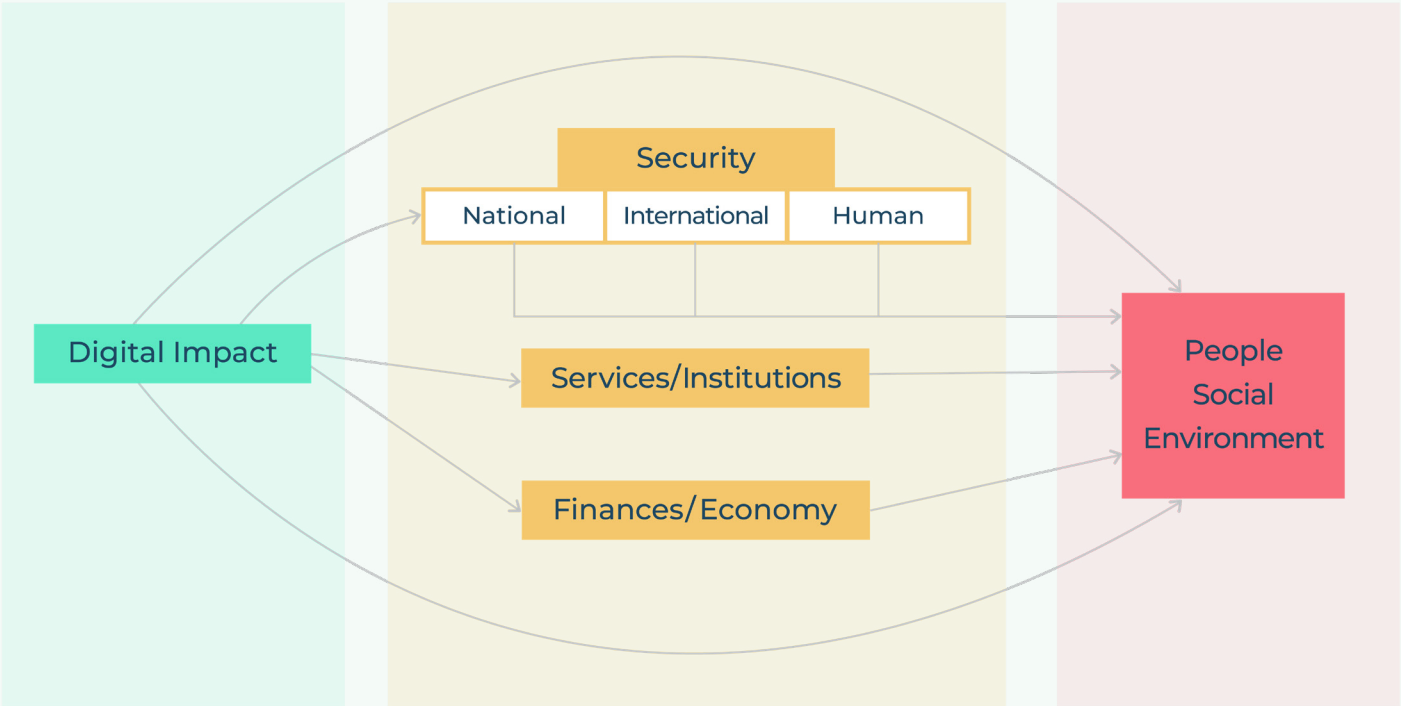
A negative impact on the victim or victims' physical, psychological, social, economic well-being, their physical security, their economic security, or on the environment.

The CyberPeace Institute proposes the following Typology of Harms caused by cyberattacks or cyber incidents. A cyberattack is defined as:

- **An attack conducted by a threat actor using a computer network or system with the intention to disrupt, disable, destroy, control, manipulate, or surveil a computing environment/ infrastructure and/or data, or to influence targets' perceptions or behaviour.**

The first indicator of an attack may appear anywhere in the typology below.

The initial impact (direct impact) must come from digital means or be at a digital level. This can then have an additional or spillover impact (indirect impact) on people or the environment. Alternatively - and much more likely, - the impact on people, society or the environment is caused by or resulting from digital means or the digital impact on security, services, institutions, finances, or the economy.



The potential impact of a cyberattack or incident may be wide-ranging, it includes:

- impact at an individual, group, company, community or country or international level;
- a threat to international, national or human security;
- limiting or depriving access to the internet;
- material or economic damage;
- deprivation of services (health, education, income, etc.,), resources (water, food, etc.) and income and employment.

A clear understanding of the impact of a cyberattack or incident would indicate:

- whether there has been an increased vulnerability of individuals or communities affected by the attack, and/or whether the attack has produced effects that are indiscriminate and cannot be controlled;
- the link between the technical impact (on computers or computer systems, e.g. the loss of functionality, whether temporary or permanent, reversible or irreversible), escalating or one-off attacks, and the resulting impact on people's lives and well-being or on the environment, e.g. tampering or deleting essential data shutting down vital services;
- an accumulation of attacks, rather than one single attack, could exacerbate each other and magnify the attack's severity. A cyberattack carried out concurrently with kinetic attacks could also amplify the ultimate harm.
- the kind of attack or incident that has taken place (ransomware, hacktivism, hack & leak, malware, DDoS, cyber influence operations, deprivation of access to internet, etc.) indicating intent which has a link to capacity for cyberattack;
- the impact is also determined by the persistent nature and scale of the attack based on number of victims (direct and indirect), the extent of damage, geographical spread, duration and temporal nature;
- the timing and location of an attack has a bearing on impact e.g. a cyberattack on a power grid during winter increases the potential of a harmful effect, or if carried out during an armed conflict where the attack adds an additional layer of harm, or the attack is synchronized with kinetic attacks or attacks across multiple sites;
- the motivation or intent of the perpetrator, e.g., financial, ideological, coercion or ego, as well as the manner of the commission of the attack. For example, whether carried out as part of a plan or policy, e.g. to exploit the information space to sow discord, or spread fear and anxiety, or to target a civilian object to impact a population rather than targeting a legitimate military objective, or to be intentionally indiscriminate which may be evidenced from the type of attack (i.e. wiper malware) or techniques used, and whether precautions were taken in crafting or deploying the attack to avoid or minimize harm to civilians;

- where, how and why individual computers or computer systems are vulnerable to ICT incidents. For example, cyberattacks on industrial control systems (water treatment facilities, dams, hospitals, etc.) on which society depends and which underpin the functioning of society would have particularly devastating physical effects.

The context in which a cyber operation or attack takes place also is important, e.g. during peacetime or an armed conflict.

For example, during armed conflict harm to civilians must be avoided to the greatest extent possible. With the use of cyber operations, the challenge is that it may be extremely difficult[26] to assess potential harm to civilians and hence to do the calculation as to whether the harm caused to civilians is proportionate or excessive to the military objective.

The definition of a cyberattack and Typology of Harm help to qualify the impact of cyberattacks and incidents and the associated risk factors.

The level of risk is a function of the likelihood/probability - which depends on intent, resources, vulnerability - that a harmful event will occur and of the severity of the impact - which must be understood - if an event does occur. Measuring harm and measuring risk are both important and complementary approaches. Harm does not happen in a vacuum. Thus, it will be important to assess risk, and the Institute will review how this work on a harms methodology can be leveraged to mitigate risk.

A recent taxonomy of the United Nations Institute for Disarmament Research (UNIDIR) addressing Artificial Intelligence (AI) risks on international peace and security[27], also emphasizes the significance of a comprehensive framework to comprehend interconnected risks and mitigate these risks and govern the development and use of AI.

The mapping of risk is correlated to an analysis of the probability of an effect or consequence. In particular how AI could heighten the risks of armed conflict or contribute to a variety of adverse effects on international security. These effects include the potential for accidents and both intended and inadvertent escalation in armed conflict, posing significant challenges for states to manage other undesirable consequences (e.g., the use of specific weapon systems), or leading to heightened tensions among states and a deterioration of regional and multilateral relations. Here, risk is analyzed according to the definitions of the International Organization for Standardization (ISO): the *"effect of uncertainty on objectives,"* and the National Institute of Standards and Technology (NIST) of the United States Department of Commerce: a *"composite measure of an event's probability of occurring and the magnitude or degree of the consequences of the corresponding event."*

For mitigation, prevention and accountability in relation to harms, a key challenge is how the impact can be quantified. It is important to refer to the proposed definitions outlined in the Lexicon and the explanations as to how the Institute relates impact to, and is distinct from harm.

Methodologies and Categories of harm

As part of its research to develop a methodology for analyzing different types of harm of cyberattacks and incidents, the CyberPeace Institute examined several different perspectives on assessing cyber harms put forward by Ioannis Agrofiotis et al.[28] (University of Oxford), CyberGreen Institute[29] [30], Deloitte[31], ICRC[32], Shandler et al.[33], UNIDIR[34], Shires & Egloff[35]. The CyberPeace Institute also commissioned research (unpublished) on measuring harm.

These methodologies demonstrate a range of different approaches to categorization, data collection, and overall assessment or rating of harms related to the malicious use of cyber. A full report has been developed analyzing these methodologies, the following is a summary of our observations.

These works can be split into two main approaches to categorizing harms related to the malicious use of cyber, with neither approach to categorizing precluding the other:

1. Separate harms in terms of their context (e.g sector-specific harm, attack specific, human impact), or
2. Describe harms in terms of their specific type.

These include all or some combination of the following harms: Physical, Psychological, Economic, Political, Reputational, Social.

Research on harms also highlights other categories including: Digital (damage to digital infrastructure, blocking access to data, disruption to media platforms); Intersectional (disproportionate effects on individuals or groups based on social categorizations); International Security; Cultural Identity; Environmental.

The Institute found it prudent to simplify these to the following categories which encompass the holistic range of harms concisely, and in order to enable meaningful measurement:

I. Physical

II. Psychological

III. Social

IV. Deprivational

V. Environmental

Much of the pre-existing literature places a high importance on both qualitative and quantitative data collection.

However, some put a stronger emphasis on quantitative data. While trying to assess all qualitative evidence of harms from cyberattacks and incidents may be unrealistic, qualitative metrics inform on quantitative data to enable a more holistic and sophisticated analysis and mapping of cyber harm.

Table Summarizing approaches of different reports:

	Categorization of Harms	Data Collected	Scoring Methods
Agrofiotis et al.	Type	Qual & Quant	N/A
UNIDIR Taxonomy	Type	Qual & Quant	N/A
ICRC	Context	Qual & Quant	N/A
CyberPeace Institute	Context	Qual & Quant	Abstract
Deloitte	Type	Quant	Self-explanatory
Shandler et al.	Type	Qual & Quant	N/A

Egloff and Shires research related to offensive cyber capabilities (OCCs) and state violence proposes an expanded definition of violence as international proximate harm to areas of human value - including the body, affective life (includes psychological or emotional harm), and social relationships. Affective life, rests at the level of the individual, and community, captures the value of relations between individuals. Harm to one can cascade into others. It is also anthropocentric, as it does not include damage to robots, animals and ecosystems unless such damage affects humans in some way. It does not include damage to property or infrastructure unless such damage affects the areas of human value. Violent acts must be intended to cause harm.

The safeguarding of data **Confidentiality-Integrity-Availability** - often referred to as the CIA triad - provides a framework to analyze and quantify the assessment of threats by cybersecurity teams. It involves the implementation of risk management procedures which include identifying information and related assets, potential threats, vulnerabilities, evaluating risks and determining how to address them. The CIA triad poses several important questions regarding the measurement of harm.[36]

- If confidentiality, integrity, and/or accountability were affected by a cyberattack or incident, what would that mean in terms of assessing the degree of societal harm caused by a cyberattack?
- Is it possible to classify the degree of damage based on the impact on digital assets? This presents an additional difficulty: how can we determine the level of the harm precisely if we do not fully comprehend the assets involved?
- The standard risk assessment methodology used considers impact in terms of revenue loss, damage to physical assets, and length of disruption. Using these impact categories alone omits consideration of the human impact.
- Assessing harm may entail identifying assets or potential targets of threat sources, such as information resources, (e.g., information, data repositories, information systems, applications, information technologies, communications links), people, and physical resources, (e.g., buildings, power supplies) that could be impacted by

cyberattacks[37] or incidents. This would enable a broad assessment of the impact of cyber threats and contribute to the goal of understanding how the loss of CIA impacts society on many levels.

An assessment of harm would entail examining the specific pillars of information security that have been affected and the way they have been impacted. This observation further underscores the importance of understanding the immediate impacts in the process of understanding secondary harms or effects, and ultimately the negative impacts on people.

As an illustration, a breach of confidentiality would entail unauthorized access to sensitive information. The extent of harm resulting from such a breach is contingent upon the level of sensitivity of the compromised data and the potential consequences from its exposure. A breach of data integrity has the potential to result in the dissemination of false or altered information. The extent of harm caused may be assessed by considering the criticality of the compromised data and the possible consequences of relying on inaccurate data.

A cyberattack or incident that impairs the availability of a system has a detrimental impact on an organization's operational capacity and its ability to provide services to its customers, beneficiaries or constituents. Digital harm has been measured by the duration and impact of the downtime of the systems, and the criticality of the affected services for a population which may help determine the extent of human or societal harm.

Assessing these approaches highlighted four key takeaways for developing a comprehensive harms methodology.

1. **A methodology should be:**
   a. adaptable to new developments and lessons learned;
   b. developed alongside further research into the mechanisms and definitions of harms from cyber.[38]

2. **Metrics should be chosen with a consideration of the context of the attack or operation, (target, method, industry, etc.).**

3. **A mature methodology will include specific guidance on data collection. A lack of standardization of data collection methods may undermine the reliability of the methodology.**

4. **A holistic methodology faces inherent challenges if it aims to measure different kinds of harm under one standard, e.g. summarizing both physical and psychological harm with one figure.**

Measuring harm

Of the different perspectives on assessing harms reviewed by the Institute, only two approaches (Deloitte, CyberPeace Institute) mention explicit scoring methodologies. These methodologies can be divided into 2 categories:

1. Abstract scoring methods, which aggregate a range of harms and weights in order to score an operation with an abstracted number that has meaning relative to the scores of other attacks.

2. Self-explanatory scores numbers that directly represent magnitude of harm and need little explanation (number of deaths, cost in US $).

Thus, it was posited that a holistic methodology faces inherent challenges if it aims to measure different kinds of harm under one standard, e.g., summarizing both physical and psychological harm with one figure.

One possible scoring methodology that avoids these complications would be to present four different assessments, each through the lens of a different type of harm. The result would be a score card of four separate ratings, one for each of the physical, psychological, social and deprivational harms.

This approach would avoid the problem of having to equate different kinds of harm, since harms within one category should by definition be commensurable, and therefore far more easily aggregated. This may also further compartmentalize the exercise of weighting the different values generated by indicators, automatically providing more granular guidance on data collection. These scores could be a combination of abstract and self-explanatory scores, such as cost in dollars, or psychological harm from a scale of 1 to 10, or on a scale from “negligible to severe”.

This still leaves the significant challenges of deciding what metrics are sufficient for assessing a given harm, creating standardized methodologies for data collection, and ensuring that harms are commensurable, even among harms of the same type.

Inevitably, any measurement of harms caused by cyberattacks and incidents is complicated by access to and reliability of pertinent data together with measurement disparities and biases. In addition, whether harm from a cyberattack or incident is direct or indirect may be far from clear.

Determining indicators of harm: **Case studies**

The Institute developed and presented two case studies based on public data on real cyberattacks, aggregating indicators of harm according to different categories determined by available data.

This demonstrated the relevance of case specific harms and general categories of harm such as physical or psychological harm. A visualization of these harm categories and indicators was made of each case study extrapolated from a large spreadsheet of indicators. For simplicity, the Institute did not endeavor to link harm categories. A number of questions on these indicators were asked to the Experts with a view to gaining insights. For a full summary of the case studies, see Annex 2.

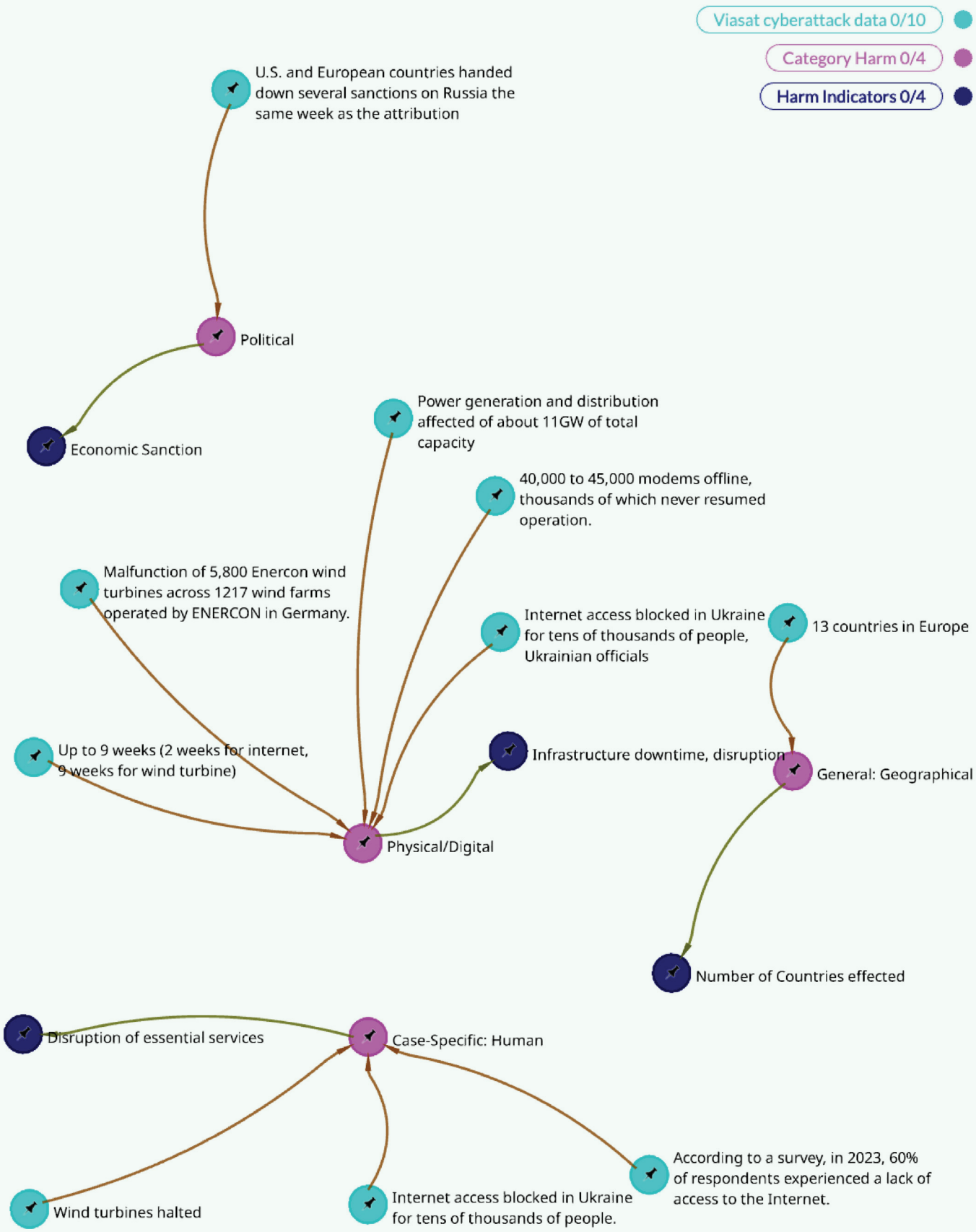
Case Study 1:

The Institute identified a range of indicators of impact and indicators from research of publicly available data on the ViaSat[39] cyberattack which took place in February 2022 in the context of the international armed conflict between the Russian Federation and Ukraine.

A cyberattack (using wiper malware) disrupted broadband satellite internet access, disabling modems which supply internet access to tens of thousands of people in Ukraine and Europe.

We identified indicators including:

1. Contextual Indicators:
   a. Geographical impact
      i. Satellite providers in Ukraine and across Europe were impacted
   b. Social impact
      i. Civilians experienced internet outages and disruptions to energy systems
2. Case-specific indicators:
   a. Operational Impact
      i. 40,000 to 45,000 modems offline, thousands of which never resumed operation.
      ii. Remote monitoring and control of 5,800 wind turbines across 1,217 wind farms.
      iii. The recovery time varied, though some were without internet for two weeks and for the wind turbine to be back online it took about 9 weeks.
   b. Human Impact
      i. Primarily, the attack impacted tens of thousands of Ukrainian civilian population as they were not able to access reliable information from the government during the conflict.
      ii. Secondarily, civilians in other EU countries experienced internet outage due to the spillover effect of the attack.
   c. Political Impact
      i. The U.S. and European countries handed down several sanctions on Russia the same week as the attribution was made in May 2022.

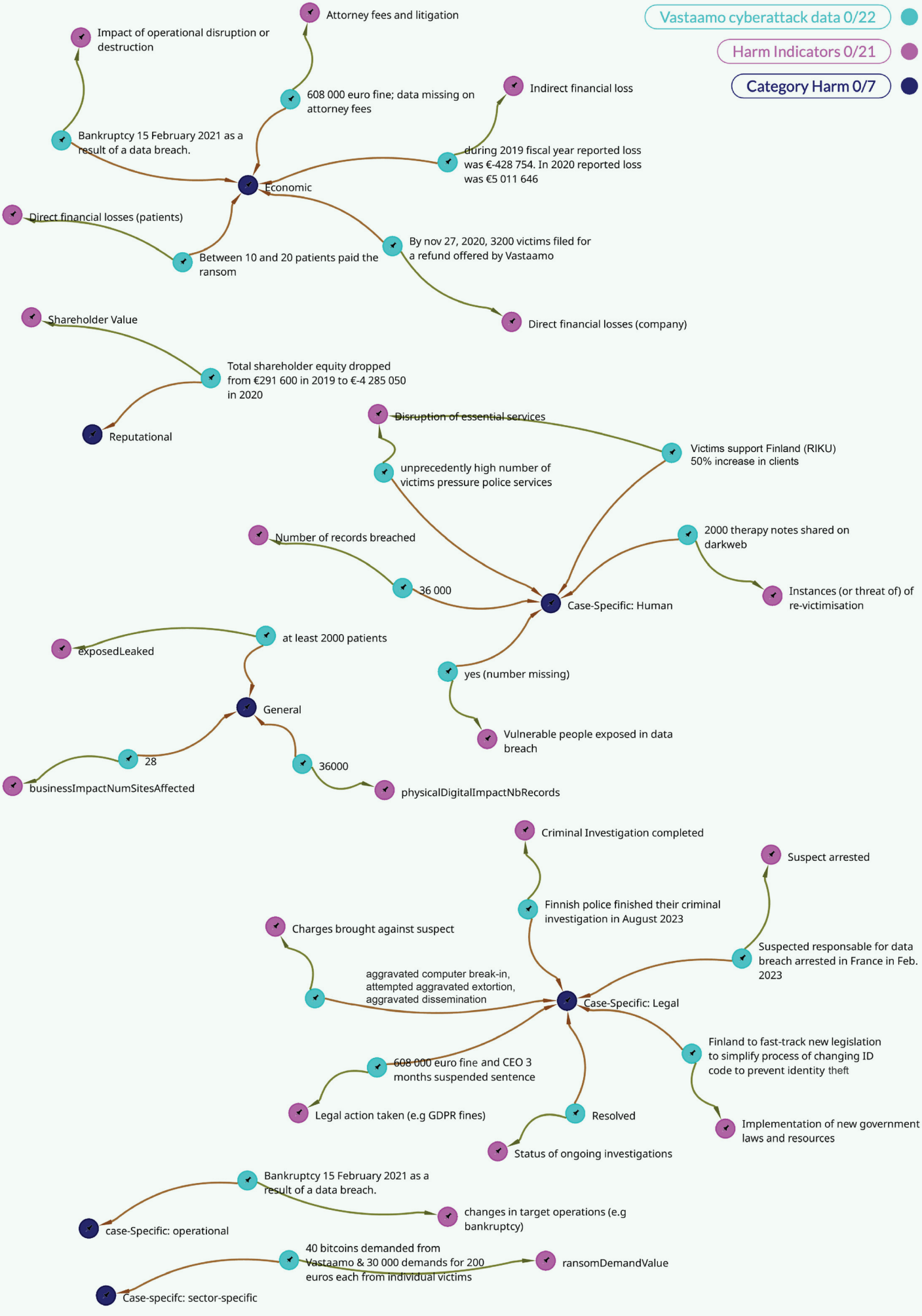

Viasat cyberattack data 0/10
Category Harm 0/4
Harm Indicators 0/4
U.S. and European countries handed down several sanctions on Russia the same week as the attribution
Political
Economic Sanction
Power generation and distribution affected of about 11GW of total capacity
40,000 to 45,000 modems offline, thousands of which never resumed operation.
Malfunction of 5,800 Enercon wind turbines across 1217 wind farms operated by ENERCON in Germany.
Internet access blocked in Ukraine for tens of thousands of people, Ukrainian officials
13 countries in Europe
Up to 9 weeks (2 weeks for internet, 9 weeks for wind turbine)
Infrastructure downtime, disruption
General: Geographical
Physical/Digital
Number of Countries effected
Disruption of essential services
Case-Specific: Human
According to a survey, in 2023, 60% of respondents experienced a lack of access to the Internet.
Wind turbines halted
Internet access blocked in Ukraine for tens of thousands of people.

Case Study 2:

The Institute identified a range of indicators of impact from research of publicly available data on the cyberattack against Vastaamo, a Helsinki-based private psychotherapy center providing private mental-health services to its patients. In late September 2020, the Vastaamo Psychotherapy Center was made aware that its systems were breached on two separate occasions in November 2018 and March 2019. A ransom payment was demanded, which when refused led to the attackers posting batches of patient records on underground forums and requesting that patients pay to have their information taken offline.

The Institute identified a range of indicators of impact from research of publicly available data, as follows:

1. Contextual Indicators:
   a. Geographical impact
      i. Within one country - Finland
   b. Social impact
      i. Impact on progress made on removing stigma around mental health
      ii. Access to essential services impacted
2. Case-specific indicators:
   a. Operational Impact
      i. 28 premises across the country were impacted
   b. Socio-economic Impact
      i. Vastaamo has since filed for bankruptcy
   c. Human Impact
      i. 36,000 patient records breached, 25,000 cases reported to police
      ii. The hack targeted vulnerable people including children
      iii. Psychological impact of the hack led to overwhelming of mental health and victim support charities
      iv. Potential for re-victimization of patients
   d. Legal Impact
      i. Criminal investigation has concluded in Finland with suspect Julius Kivimäki on remand since February 2023
      ii. Vastaamo was found to have infringed GDPR and was issued an administrative fine and reprimand
      iii. Former Vastaamo CEO, Ville Tapio, given a three-month suspended sentence for failing to meet obligations under GDPR.


Vastaamo cyberattack data 0/22
Harm Indicators 0/21
Category Harm 0/7
Attorney fees and litigation
Impact of operational disruption or destruction
608 000 euro fine; data missing on attorney fees
Indirect financial loss
Bankruptcy 15 February 2021 as a result of a data breach.
during 2019 fiscal year reported loss was €-428 754. In 2020 reported loss was €5 011 646
Economic
Direct financial losses (patients)
By nov 27, 2020, 3200 victims filed for a refund offered by Vastaamo
Between 10 and 20 patients paid the ransom
Direct financial losses (company)
Shareholder Value
Total shareholder equity dropped from €291 600 in 2019 to €-4 285 050 in 2020
Reputational
Disruption of essential services
Victims support Finland (RIKU) 50% increase in clients
unprecedently high number of victims pressure police services
Number of records breached
2000 therapy notes shared on darkweb
36 000
Case-Specific: Human
Instances (or threat of) of re-victimisation
at least 2000 patients
exposedLeaked
yes (number missing)
General
Vulnerable people exposed in data breach
28
36000
businessImpactNumSitesAffected
physicalDigitalImpactNbRecords
Criminal Investigation completed
Suspect arrested
Finnish police finished their criminal investigation in August 2023
Charges brought against suspect
Suspected responsable for data breach arrested in France in Feb. 2023
aggravated computer break-in, attempted aggravated extortion, aggravated dissemination
Case-Specific: Legal
Finland to fast-track new legislation to simplify process of changing ID code to prevent identity theft
608 000 euro fine and CEO 3 months suspended sentence
Legal action taken (e.g GDPR fines)
Resolved
Implementation of new government laws and resources
Status of ongoing investigations
Bankruptcy 15 February 2021 as a result of a data breach.
case-Specific: operational
changes in target operations (e.g bankruptcy)
40 bitcoins demanded from Vastaamo & 30 000 demands for 200 euros each from individual victims
ransomDemandValue
Case-specifc: sector-specific

Lexicon

The CyberPeace Institute developed a Lexicon for the Harms Methodology according to the following approach:

i) Extracted specific key terms and phrases that were the subject of discussion at the Experts Meeting, and/ or had been terms that had already been worked on or developed ahead of the Meeting.

ii) Identified fundamental tensions within definitions, with reference to other relevant published definitions, e.g ICRC and Council of Europe Committee.

iii) From these tensions formulated a series of questions for a brainstorming session.

iv) Repeated steps ii & iii iteratively to produce the current version of the lexicon.

This Lexicon will be the subject of consultations going forward.

1. Cyberattack

An attack conducted by a threat actor using a computer network or system with the intention to disrupt, disable, destroy, control, manipulate, or surveil a computing environment/infrastructure and/or data or to influence targets' perceptions or behaviour.

With reference to:

- Council of Europe, Cybercrime Prevention Committee
- Tallinn Manual
- State positions on definition of cyberattack

Note:

The current CyberPeace Institute definition of cyberattack. We note that various states have different interpretations of how the use of cyber operations would meet the threshold of an armed attack according to International Humanitarian law.

2. Cyber violence

The purposeful use, threat of use, negligent use or autonomous action of digital and information technologies that directly, indirectly, temporarily or permanently causes either immediate or long-term harm, determined as negative impact on people's health, their physical security, their economic security or the environment.

With reference to:

- Council of Europe, Cybercrime Prevention Committee
- WHO definition of violence
- WHO definition of health
- CyberPeace Institute comparative analysis of pre-existing work on harms of cyber incidents
- CyberPeace Institute Expert Meeting feedback

Note:

This definition is one of the core contributions of the CyberPeace Institute. It is derived in part from the WHO definition of violence, and assumes the WHO definition of health which is "A state of complete physical, mental and social well being." The Cyber violence definition also considers the Council

of Europe Cybercrime Prevention Committee definition of cyber violence, and recognises that both definitions are largely compatible. However, the CyberPeace Institute definition is more closely harmonized with the categories of harm and other established terms that are referred to in the Harms Methodology.

Several iterations of this definition have been socialized and revised in consultation with internal and multistakeholder experts, namely within the CyberPeace Institute's Expert Meeting in November 2023.

The question of including "institutions" in the definition was considered in our framing. However, it was determined not to include institutions in this definition, recognizing that harm from attacks and incidents that affect "institutions" would be included in the data collection to assess its harm on people, society and the environment.

3. Direct result

Broadly refers to results that would be impossible without the given incident.

With reference to:

- CyberPeace Institute Expert Meeting feedback
- State positions

Note:

Defined in order to contrast and provide context for the term "Indirect result." The key difference is the certainty with which an incident can be said to have caused a result.

4. Harm[40]

A negative impact on the victim or victims' physical, psychological, social, economic well-being, their physical security, their economic security, or on the environment.

With reference to:

- WHO definition of health
- WHO definition of violence
- Coupland, Taback, Dobos' Theory of violence
- CyberPeace Institute comparative analysis of pre-existing work on harms of cyber incidents
- CyberPeace Institute expert meeting feedback

Note:

The definition is inferred from the WHO definition of violence and assumes that the possible negative impacts of violence are synonymous with harm. The following four categories of harm were synthesized from the WHO definition of violence and an analysis of pre-existing work on assessing harms of cyber incidents. These categories were then stress-tested during the CyberPeace Institute's Expert Meeting, and adapted as per the feedback provided.

A single incident may cause multiple kinds of harm that should be considered separately, (e.g the physical and psychological impacts of a single act of violence), and many different harms may be causally interconnected. We understand the range of harms below to be applicable to victims including individuals and groups, and those affected when an institution is targeted.

a. Physical harm

Injury or death to persons or damage or destruction to physical or digital objects.

With reference to:

- Tallinn Manual
- CyberPeace Institute comparative analysis of pre-existing work on harms of cyber incidents.
- CyberPeace Institute Expert Meeting feedback
- State positions

Note:

This definition is derived from the Tallinn Manual, with the addition of "digital" to include harm to digital infrastructure. This addition was informed by the CyberPeace comparative analysis and feedback from the Expert Meeting. It should be noted that there is currently no agreement on whether digital data is an object under IHL.

b. Psychological harm

Severe mental suffering that is tantamount to injury.

With reference to:

- Tallinn Manual
- CyberPeace Institute comparative analysis of pre-existing work on harms of cyber incidents.
- CyberPeace Institute Expert Meeting feedback

Note:

This definition was taken directly from the Tallinn Manual. Experts have expressed concerns around prohibitive factors for data collection around psychological harm, and suggested the methodology draws on preexisting psychiatric measures of psychological harm within the insurance industry and instances of national law.

c. Social harm

Refers to damage that affects autonomy, development and growth, and access to cultural, intellectual, informational resources.

With reference to:

- UNIDIR Taxonomy of cyber harm
- WHO definition of health
- CyberPeace Institute comparative analysis of pre-existing work on harms of cyber incidents.
- CyberPeace Institute Expert Meeting feedback
- Academic research

Note:

Definition taken from UNIDIR Taxonomy of cyber harm, where it is defined as "cultural harm" and applies specifically to societies. We have referred to these harms as "social harm" to harmonize this concept with the WHO definition of health, which mentions social well-being. We further expand on the UNIDIR definition to include individuals, groups and institutions, so as to encompass harms resulting from mis/disinformation, undermined trust in institutions, and harms unique to political institutions. Experts noted that these harms must specifically be covered by a comprehensive harms assessment.

Experts were unclear as to whether the term "social" or "societal" should be applied here. While there are no unanimously accepted explicit definitions of these terms, we understand "societal" as placing emphasis on societies and institutions as entities that are separate to the individuals that constitute them.

Consequently, we understand societal harm as encompassing negative impacts on an institution's ability to function as intended, with less of an emphasis on specific human impacts. Contrastingly, we understand "social" as placing emphasis on the social experience of humans within a community or institution, and "social harm" to encompass negative impacts on the social experiences of people. Our current decision to apply the term "social harm" is determined by the scope of our methodology, which considers specifically harms to humans and the environment, recognizing that harm from incidents that affect institutions would be included in the data collection to assess its harm on people and the environment.

Future iterations of this definition will be further informed by developments around data collection strategies which will establish how the different categories map to accessible data.

d. Environmental harm

Widespread, long-term and severe damage to the natural environment.

With reference to:

- International Humanitarian Law
- CyberPeace Institute comparative analysis of pre-existing work on harms of cyber incidents.
- CyberPeace Institute Expert Meeting feedback

Note:

This definition is taken directly from customary international humanitarian law. We currently do not suggest indicators for this harm, however it was found to be an important feature for futureproofing the definition of harm.

e. Deprivational harm

Acts causing material or economic damage, depriving of resources, loss affecting victim's access to essential services, income, employment, education, skills, and living/working environment.

With reference to:

- CyberPeace Institute comparative analysis of pre-existing work on harms of cyber incidents.
- CyberPeace Institute Expert Meeting feedback

Note: Deprivation includes reputational harms.

5. Indirect result

Broadly refers to results that may have been impossible or less severe without the given incident. This includes cascading results that are caused by the direct results of an incident, (also referred to as secondary effects) and the results of operations that combine cyber and other elements (including instances of cyber-enabled crime).

With reference To:

- UN Ad Hoc Committee on Cybercrime
- CyberPeace Institute Expert Meeting feedback
- State positions

Note:

The CyberPeace Institute recognises that quantifying all the indirect harms of a given incident is not feasible. However,

many states acknowledge that indirect impacts of cyber incidents are significant.

Therefore, this broad definition aims to meet this need without setting an impossibly high standard for assessing harms. It gives space for a holistic assessment of harms, but the degree to which indirect results are considered should be determined on a case-by-case basis by the harms methodology.

6. Theory of Violence

The Theory of Violence assumes that violence can always be expressed in terms of its impact on the victim's health, e.g. lethality, number of people killed, injured, displaced, assaulted, etc. The determinants of the impact of any act of violence in any context are:

a. Intent

The intent of the perpetrator to cause the impact.

b. Physical Capacity

The physical capacity of the perpetrator for violence (given by the number of guns, knives, etc. available to cause the impact in question).

c. Capacity for a Cyberattack

The capacity of the computer systems available to the perpetrator and the technical expertise to launch an attack.

d. Vulnerability

The vulnerability of the victim or victims (given by the potential of the victim to suffer the impact in question).

e. Impact

A broad term describing the effects of an incident, including human harms and effects on security, services / institutions or finances / the economy. The potential human aspect of the impact of a cyberattack or incident is wide-ranging.

It includes:

- Impact at an individual, group, company, community or country level;
- A threat to international, national or human security;
- Limiting or depriving access to the internet;
- Material or economic damage;
- Deprivation of services (health, education etc.,) access to work and resources (water, food etc.;).

With reference To:

- Coupland, Taback, Dobos' Theory of Violence
- Typology of Cyber Harm
- WHO definition of violence.

7. Victims

Persons who, individually or collectively, have suffered direct or indirect harm.

With reference to:

- International Criminal Court Rules of Procedure and Evidence
- Declaration of Basic Principles of Justice For Victims of Crime and Abuse of Power[41]
- Basic Principles and Guidelines on the Right to a Remedy and Reparation for Victims of Gross Violations of International Human Rights Law and Serious Violations of International Humanitarian Law[42]
- WHO definition of harm

- CyberPeace Institute Expert Meeting feedback
- CyberPeace Institute comparative analysis of pre-existing work on harms of cyber incidents.

Note:

This definition is adapted from the definition presented in the Declaration of Basic Principles of Justice for Victims of Crime and Abuse of Power. Changes have been made to harmonize it with other terms presented here for the purposes of the Harms Methodology. These changes include:

- Removing the qualifications of harm, and instead taking the definition of harm presented in this Lexicon. We understand the UN General Assembly's qualification of harm that "includes physical or mental injury, emotional suffering, economic loss or substantial impairment of their fundamental rights," is encompassed by the definition of harm presented in this Lexicon.
- Adding the concept of indirect harms.
- Removing legal specifications to make the definition legally agnostic.
- We acknowledge that different legal contexts define the term "victims" differently (see Notion of harm and victim in international law, p48). One significant point of divergence is in whether to consider institutions as potential victims, as in the ICC definition.

  For the purposes of this Methodology we are concerned with harm to humans and the environment. Therefore, we are concerned with harm to institutions insofar as it leads to human harms, and have focused our definition of "victims" not "persons".

8. Violence

The intentional use or threat of physical force against a person, group or community, that has a negative impact on the victims' physical, psychological or social wellbeing including deprivation.

With reference to:

- WHO definition of violence.

Note:

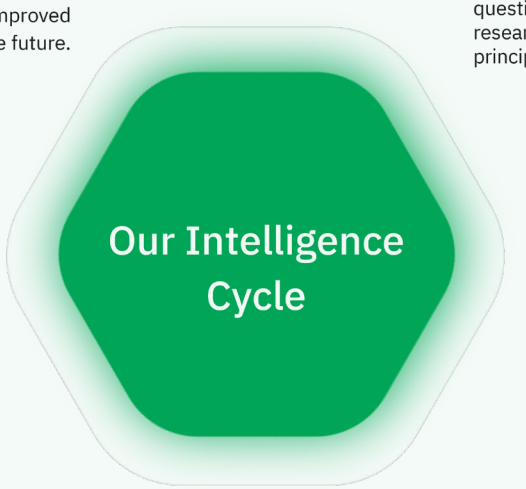
Directly derived from WHO definition: "The intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community, that either results in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment, or deprivation." This definition includes key language that is already widely accepted by states.

Data collection model

Data is the backbone of the work of the CyberPeace Institute, leveraging and processing data from our partners, open sources, proprietary sources and that produced through our own collection processes. Our data pipeline processes data from its raw form to information suitable for analysis. It capitalizes on the use of scalable cloud-based technologies. Our processes put data privacy and security at the forefront of decision-making and we build increasingly automated data flows to reduce processing errors, increase data recency and reduce time to analysis.

Different data-centric projects call for different analytical approaches. Our portfolio of analysis activities includes but is not limited to problem profiling, Open Source Intelligence (OSINT), Social Network Analysis, technical analysis including log, malware and forensic analysis, Geotemporal analysis, financial asset tracing and victim profiling. We also have a network of trusted partners, whose expertise we can call-upon for other specific types of analysis.

The Institute has extensive experience in conducting research to identify the individuals or groups responsible for cyberattacks and incidents that target vulnerable communities. Our research also aims to uncover the motives behind these attacks and assess the harm they inflict. We will leverage our existing analytical and intelligence methods, along with our data collection model as shown in the graph below, to effectively implement the process of mapping harm.



Feedback

we learn from every project we deliver on. Taking on feedback from our community, our governance bodies and our partners we strive to produce improved analytical products in the future.

Direction

every research project begins with the setting of clear intelligence requirements and the definition of research questions that need to be answered. This ensures our research stays within scope, respects ethical research principles and avoids mission creep.

Disseminate & Share

complex analysis must be accompanied by simple storytelling. Developing data-visualization platforms tailored to each research project and publishing clear reports and infographics allow us to communicate our findings with our community.

Data collection

manually, automatically or somewhere in between, we collect data from primary data sources, open sources and closed sources which when combined together gives us a more comprehensive understanding of the cyber risks faced by vulnerable communities.

Analysis

from data discovery, statistical analysis, Social Network Analysis to geotemporal analysis, we find hidden connections within large datasets. Using data visualization and analysis tools, including dashboards and graphical link analysis software, our analysts can connect information from disparate sources to find the answers to our research questions.

Processing

using data pipelines we clean and normalize data and evaluate its relevance and reliability to transform it from its raw form to exploitable information usable for analysis. This step of our intelligence cycle requires close collaboration between our analysts and technical engineers.

The Institute follows a structured approach to its ongoing data collection and analysis, which for the work on the Harm Methodology we are refining through the development of case studies, of which two preliminary case studies are included in this Report and were presented to the Experts Meeting. We use a systematic spreadsheet format for recording data. Primary fields in this collection include the categories of harm, indicators, data availability, impact details and relevant source information. This approach not only facilitates ease of analysis but also ensures the data can be converted into a database format for enhanced management and scalability.

The Institute uses a variety of methods for its analysis and data visualization phase to methodically interpret the datasets. By applying network and link analysis, the analysis can identify patterns and correlations within the data. Utilizing tools like GraphXR, the datasets can be visualized in a comprehensible manner, facilitating the extraction of actionable insights, and contributing to a deeper understanding of the harm caused by cyberattacks.

As a further step, the Institute is actively researching and developing its machine learning capabilities to process and collect information from unstructured datasets. The initiative aims to test automation of the data collection process, potentially enhancing accuracy and speed, as well as scale.

The Institute will continue to refine the Lexicon mentioned earlier through ongoing research and consultations with experts in the coming months, as this Lexicon will be important to frame data collection. Furthermore, the examination of additional case studies could offer novel perspectives on the ongoing research conducted in this study regarding the categories and indicators of harm for the data collection. Thus, to expand our analysis we will continue to work on case studies, and are currently assessing capabilities of artificial intelligence (AI) to monitor and evaluate the impacts of a cyberattack or incident.

Observations from the Expert Meeting

The format of the Expert Meeting was presentations and background documents provided to participants in a series of plenary sessions followed by discussions in working groups, which were captured and written up by staff from the Institute. The following is a summary of the feedback and insights provided by experts during the meeting, grouped according to the specific themes addressed in guiding questions or as determined by the experts feedback.

This feedback is followed by clarification or how the feedback is addressed by the Institute, as highlighted in the text.

On Purpose of the Methodology

At the opening of the Expert Meeting, questions were asked by participants with regard to the purpose of the Harms Methodology, particularly whether it aims to be a legal tool.

- As highlighted in the introduction, the objective is to determine the means to measure harm from cyberattacks and incidents in order to increase knowledge of the human costs, and influence policy, accountability and resilience efforts.
- The Institute does not aim for this Methodology to be a legal tool but to support ongoing accountability efforts. The Institute does not conduct its own attribution[43] of cyber incidents to identify the actor(s) involved but documents the attribution efforts by others to link a particular individual, group or state to a specific incident.[44]

It is for States to determine what cyber operations they consider as a violation, if attributable to another State, and this is what States have been asked to do as part of the UN OEWG.

The Institute was advised by Experts to ensure the Methodology uses defined and understandable language, not be overly scientific or complicated in nature and to ensure that this did not become a purely academic exercise.

Experts noted that when discussing quantified harm, it appears to speak to all stakeholders initially. However, they do not seem to be actually acting on the harms in the longer term. It was noted that some cyberattacks and incidents have the same methods but ultimately have different levels of impact, and this would be important for policy makers to understand. The Harms Methodology could thus be relevant in this regard.

- The Institute determined the relevance of continuing with the development of the Harms Methodology.

In introducing this work at the Expert Meeting, feedback highlighted the importance of carrying out advocacy efforts in parallel to developing a Harms Methodology to both socialize and explain the methodology and approach. This would require specific strategies targeting different stakeholder groups. The narrative could reference behaviors that need to be encouraged and discouraged, and place an emphasis on vulnerability and possible preventive measures.

- For advocacy to be effective, and consistent with the Institute's evidence and data driven approach, the data to policy process requires investment in establishing facts from the data, including definitions, methodology, clear narrative, a lexicon, data collection procedures, etc. Once this is completed, a process of engagement with key stakeholders needs to be elaborated and implemented. The narrative needs to be compelling and accessible demonstrating why it is a public concern.

On the Theory of Violence

Feedback from the Experts Meeting confirmed that the Theory of Violence is applicable to cyber violence although some experts argued that inherent differences of the cyber domain may make the theory untenable underlining that not all attacks will have violent impact. Therefore, it is important to identify the relationship between violence and harm.

The Theory of Violence was viewed as important to:

- understand the dynamics of the violence, the particular vulnerability, and potentially the determination of the norm or law violated;
- diminish the likelihood of it happening again through approaches to decrease vulnerability, capacity or intent.

This is important for law enforcement, governments and targeted companies/ organizations. It can be used to contextualize neutral measures of harm. The Theory was not seen as necessary for defining harms, although some participants argued that capturing the concept of vulnerability is an important part of assessing harms.

It was advised to be cautious as to how any potential expanded definition of violence or harm or identification of harm through the Methodology may be misused by different stakeholders in the future. In particular, using the identification of, an expanded definition of, or a lower threshold of harm, to call for a retaliatory response in response to a cyberattack.[45]

- The Institute determined to maintain the Theory of Violence as an applicable reference and worked on a new definition of Cyber Violence.
- In weighing up the risks of the methodology being leveraged in this manner or manipulated, the Institute believes the benefits of a methodology outweigh the risks. The Institute will monitor the use of the methodology and definition, once finalized, to assess any positive or negative consequences of its use.
- Further consultations and research will be carried out in this regard.

The work of Egloff and Shires was useful in reaching this conclusion. This focuses on the broader definition of violence and harm and whether this could lead to a *"potential implication of conceptual expansion on (international) legal understandings of armed conflict"*.[46] They highlight the following in relation to state violence and the use of offensive cyber capabilities (OCCs)[47] :

- *"There are two major international legal frameworks that an expanded concept of violence for OCCs could affect: jus ad bellum, particularly its understanding of use of force and armed attack, and jus in bello, particularly international humanitarian law's (IHL) focuses on violence and the protection of civilians during armed conflicts. For the former, the expanded concept of violence may lead to more cyber operations being considered a use of force than a narrow conception."* Reference in this regard is made to the first Tallinn manual and the provision of 8 criteria to judge whether a cyber operation is a use of force: severity, immediacy, directness, invasiveness, measurability of effects, military character, state involvement, and presumptive legality. The first, third, and fourth criteria are perceived as potentially open to more permissive interpretations based on an expanded concept of violence.[48]

- "*Even then, an expanded concept of violence is unlikely to have any impact on the definition of 'armed attack', which is generally considered to be a higher threshold, depending on the scale and effects of the operations compared to physical precedents."* Reference is made in this regard to the Nicaragua judgment of the International Court of Justice.

On Definition of harm from malicious use of cyber

The Institute provided a draft definition of harm from the malicious use of cyber which was commented on by the experts:

- "The use of digital and information technology which results in or has a high likelihood of resulting in either directly or indirectly:
  - a negative impact on the victim or victims' physical, psychological, social or economic well-being, or on their security;
  - harmful environmental consequences."

1. Feedback was provided in relation to a specific guiding question asking how relevant "intentional use of digital technologies" was in the definition, i.e. the purpose behind the act. The feedback highlighted that:

a. malicious and deliberate wrongdoing were not relevant for the definition of violence as even acts committed in ignorance, negligently or unintentionally can cause harm and should be measured;

b. an attack could be intentional attacks but with unintentional harm;

c. there was a link between intent and vulnerability being latent and not directly measurable;

d. intent is not directly relevant to the actual harm, e.g. an individual may suffer the same injury from an accident, or rogue AI, or criminal negligence, or from a deliberate attack;

e. intent may be difficult to demonstrate empirically;

f. intent may be discovered months or more after the attack;

g. the intent of an act could also be to prevent a greater harm;

h. questions were raised about how the use of "intent" relates to legal lexicon.

Therefore, while intent is useful for measuring how an attack happened, and for attribution, prevention and other important processes, it is not relevant for developing a neutral measure for harm. However, it is important to document and measure intent for potential legal and other accountability measures.

The Institute thus amended the definition. In relation to data collection and analysis, the Institute will still collect and document intention where it is available in data captured. The Institute will need to make it explicit how the Harm Methodology deals with a range of intentional and unintentional harms.

2. There was some discussion as to whether it was necessary to include specific reference to "harmful environmental consequences" in the definition. Observations highlighted that:

a. positioning of "environmental" as a specific point calls too much attention to a highly unlikely event, e.g. a cyberattack or incident which harms the environment if the harms are from cyber means only;

b. environmental harm could be encompassed in social harms (depending on the definition);

c. it was not clear if environmental harm means harm to non-humans (e.g nature, animals), as separate from human harm (e.g even if no humans are harmed as a result);

d. environmental harm could be said to only be relevant in so far as it results in the other kinds of harm, and thus is already covered by those kinds of harm;

e. alternatively, "environmental" could be included in the definition, as the consideration of biodiversity has had more value in the normative sphere in recent years, and is linked to human well-being;

f. harm to the environment was an issue linked to the pollution caused by producing and using technology, and consuming non-renewable resources used to make technology, which was not the ambition behind this Harm Methodology.

It was determined not to remove environmental harm, or include it in social harm measured alongside the other kinds of harm but to specify clearly what this harm was related to. This is in view of the fact that environmental harm is mentioned in IHL[49].

The determination of a specific Lexicon for all terms used in the definition is essential, see section "Lexicon,"p.23.

3. It was noted that few cyberattacks and incidents actually cause physical violence, so "physical" could be confusing, however, it was recognised that it was important to reference this category.

4. It was also noted that a cyberattack or incident may have effects that are immediately observed, or there may be a lag before effects are known and thus it was necessary to monitor in a longitudinal way.

5. Inclusion of the term “victim” was discussed to clarify what victim(s) could be at individual, system, country and international levels. Feedback was provided that the term victim may be considered a legal term or related to the legal status of the person and may thus introduce unintended elements to the definition.

The proposed definition was updated and the term “victim” is no longer referenced in the definition, replaced with “people". The Lexicon makes reference to the term victim, and how this is generally defined.

6. Insights were provided with regard to whether to focus only on direct harms or whether to also include indirect harms. In this regard, concerns were raised with regard to the point at which to stop counting the indirect causes of an attack; how many degrees of separation, and over how long a time period. A stopping point must be able to be replicated across different assessments.

It was suggested that we apply an approach that is descriptive / indicative of indirect harm rather than trying to comprehensively document all actual indirect harm. Analogous to the Mercalli scale as opposed to a Richter scale for measuring impact of earthquakes.

The revised definition refers to “directly or indirectly”. In the collection of data and piloting of the Methodology, the Institute will need to define how it is assessing and documenting its definition of direct and indirect harm.

7. Scope of definition.

a. It was considered that attackers appear to be the subject of the definition, and it may make sense to have the victim as the subject.

b. It was important to review notions of reparable and irreparable harm, and replicability.

c. The original definition was criticized as being too inclusive if the word “use” was maintained. For example, it could currently include a harshly-worded email as cyber violence. The term “misuse” was suggested.

The notion of *“use”* is retained and also expanded in the new definition, as it was important to maintain an intention-agnostic measure, recognizing that it is important to note there are harms that are not helpful to consider. It is for others to assess what constitutes misuse (and other normative questions in general) based on the intention-agnostic measures of harm.

With this in mind, the way to exclude a harshly worded email from the definition of harm would not be to use the word "misuse," but rather to note that some harms are incidental and impractical and unhelpful to consider.

8. The definition requires a standard lexicon to ensure the meaning of terms is clearly understood.

a. The lexicon should harmonize concepts and definitions, (e.g. categories of harm and definition of cyber violence should list the same kinds of harm)

b. Definitions should be explicit and accessible for non-native English speakers and a range of stakeholders, including laypeople and policy makers.

The Institute proposes a Lexicon with reference to any relationship or lack thereof to legal and other technical terms that may overlap, (e.g " disclaimer: this is not a legal definition").

Further to the Expert Meeting, the Institute worked on a new definition related to Cyber Violence (which by its nature has an intent to harm) as follows:

- *"The purposeful use, threat of use, negligent use or autonomous action[50] of digital and information technologies that directly, indirectly, temporarily or permanently causes either immediate or long-term harm, determined as negative impact on people's health[51], their physical security, their economic security or the environment."*

On Typology of Harms

Experts noted the importance of understanding the dynamics of harms from the use of cyber means as moving between computers first before causing harm to people, as outlined in the Typology.

Equivalence is useful because concrete comparisons are more understandable for non-academic/scientific stakeholders. One expert suggested that when strategic outcomes of a cyber operation and a kinetic operation are the same, then those two operations could be said to be equivalent. Others disagreed, saying that the dynamics of cyber and kinetic are so inherently different, with the latter being firmly established in normative frameworks, that equating them would be difficult if not impracticable. One expert gave the example that while a bombing of a target may have a similar impact on the target as a cyberattack, the destruction, unintentional or incidental damage to persons or objects that would not be lawful military targets in the circumstances (so called "collateral damage") will play out differently.

There was some discomfort voiced concerning equivalence – on the basis that at some point the distinction between a cyberattack and kinetic attack breaks down. For example, the cutting of undersea cables and attacking of water-treatment facilities could have both physical and cyber implications. Experts, on the other hand, have acknowledged that there is a significant influence or connection between kinetic and cyber activities.

It was argued that cyberattacks and incidents often cause harm when combined with other non-cyber factors, and therefore cannot be said to cause harm to the same extent as kinetic by the following reasoning: an intervention causes an event to the extent that the event would be impossible without the intervention. Since cyber is often paired with kinetic operations, it is often unclear to what extent the impacts are caused by cyber, and to what extent they are caused by kinetic operations.

On Methodologies and Categories of Harm

A summary of the methodologies and categories of harm were presented to the Expert Meeting.

No dissenting opinions were received in relation to the four key takeaways for developing a comprehensive Harms Methodology. Advice received from the experts was that more research should be carried out with regard to how these terms are defined in other standard measures of harm.

Experts suggested that a methodology should be able to monitor and assess impact over time. Thus, it must include a temporal graph to show the causality between attacks and external factors. For example, cyber incident targeting critical infrastructure in a country after that country announced its support for sanctions being imposed on a country, or a cyberattack against the energy sector right before winter.

On the specific categories of harm, the following insights were provided:

I. Physical

- Requested clarification on how physical applies to people, and physical and digital infrastructure.

II. Psychological

- Experts suggested that psychological harm may often be caused by other kinds of harm, and so questioned the importance of giving it its own category.
- Suggested the methodology draws on existing research (including psychological expertise and insurance claims for psychological harm) for understanding of and quantifying psychological harm.
- Questions were raised about how easily data can be collected around psychological harm; there may be prohibitive factors.
- Asked for clarifications around the level of granularity at which psychological harms would be considered, e.g., specific mental illness.

III. Social/ Societal

- Questions were raised around the use of the term social or societal. (See Lexicon.)
- Lack of clarity as to what exactly was included in this harm.
- Some argued that political harm should be kept separate, given that some incidents can harm political institutions without harming people.
- Some argued that this definition does not capture broad scale disinformation campaigns that can cause loss in confidence in organizations and governments.

IV. Deprivational

- Deprivational was not understood as including economic harm on a first reading.
- Loss of digital assets / knowledge should be considered in assessing harms.
- Reference was made to reputational insurance policies as a basis for measuring reputational harm.

V. Environmental

- Likelihood of this harm was questioned.
- In relation to ICTs, generally reference is made to the impact caused by general use of technology in terms of pollution, use of non-renewable resources.

In addition to the above, the following breakdown of categories was suggested: human death, injuries, psychological impact, economic impact, political impact, reputational harm, and delayed services. It was highlighted that the methodology should include considerations of gender based violence.

Other insights were provided which have been regrouped thematically.

- **Compounding harms**

Comments were raised around how different categories of harm may be causally linked to each other and how events may cause several kinds of interrelated harm. This was found to be unproblematic in cases where the types of harm are clearly distinct. For example, being punched publicly can be clearly said to inflict compounding physical and psychological harm, but it cannot be said to inflict compounding physical harm, bodily harm, and kinetic harm, since these kinds of harms are not distinct, and so compounding these three harms could amount to counting a single punch as three punches. This also highlights the risk of overinclusive harm categories. A category that encompasses two distinct types of harm may under-represent one of them, by counting two harms as a single instance of harm.

These points highlight the importance of robust categories of harm that are clearly distinct, and not overinclusive.

- **Repairable vs. Irreparable harm**

A distinction was made between repairable and irreparable harm, as a way of classifying different degrees of harm.

- **Traditional Cybersecurity terminology**

Suggestions were made of the relevance of ensuring compatibility with the Confidentiality-Integrity-Availability (CIA) triad definition of cyber impact in the methodology.

- **Key considerations for how Categories inform the Methodology**

The 4 categories were found to satisfyingly encompass a holistic scope of human harms.

The selection of categories, defining which harms to aggregate, and which harms to quantify separately, have significant implications for the Methodology.

The categories must encompass a holistic scope of human harms, but should not group different types of harm together in a way that misrepresents harm.

It will be important to explicitly clarify categories and harmonize them with definitions. One approach could be to assemble a bulk of indicators and select the most appropriate ones within each harm category on a case by case basis. The justifications for these selections should be clearly explained.

The following was framed as a crucial research question: Do we start with the categories, and create indicators, or do we start with indicators and create categories?

Is this influenced by what data is available?

- The four key takeaways for developing a comprehensive Harms Methodology will guide its development.
- Categories of harm for the next stage of development of the Methodology will be Physical, Psychological, Social, Deprivational and Environmental. Robust categories of harm will be developed that are clearly distinct and not overly inclusive.
- More research will be carried out with regard to how categories of harm and terms are defined in other standard measures of harm. Research on the CIA triad and by Shires and Egloff has already been included herewith.
- The Institute will explicitly clarify categories and harmonize them with definitions.
- Research questions elaborated by the experts will be followed up on, particularly
  - □ Do we start with the categories, and create indicators, or do we start with indicators and create categories? Is this influenced by what data is available?
- Learnings from the aggregation of metrics for an examination of case studies will be elaborated and shared as part of the explanation of the Methodology.

On Measuring Harm

On scoring methods, the Experts Meeting provided the following observations:

- Using simple scoring methods will make advocacy easier.
- A broad tiering system was suggested, classifying attacks as minor, severe, etc.
- Harm types should be kept separate if possible, to avoid comparing the incomparable. Comparability of metrics could be made across a type of harm but not between different categories of harm, qualifying measurement as high, medium and low.
- It could be detrimental in some cases to have equivalence across the categories of harm, e.g., comparing a certain level of psychological harm against 'number of deaths' would not work. Thus, the Institute should consider indexing values for the harm levels, psychological, physical, etc.
- An abstract scoring method was criticized as having no strong basis in reality. It was argued that it would be difficult to ground an abstract score in reality, since it only has meaning with respect to other attacks, (as opposed to international law, or thresholds of attack).
- DALY and Nutriscore were given as examples for scoring metrics.
- It is possible and justifiable to quantify harm if the methods used to do so are transparent. This also allows for some degree of assumption, which can be explained to and assessed

by stakeholders like states or legal authorities.

- Regarding the weighted values that were presented, it was suggested that missing data should be clearly noted as such, rather than providing a middle weighting for those cases. Thus, it is worth researching alternative methods for weighting missing data.

The Institute will continue research into the matter of providing a scoring or weighting within the categories of harm and the validity or not of comparing harms across categories. This will be modeled using some case examples.

On Data Collection and Indicators

Experts advised that there should be separate databases of data collection in relation to effects of cyberattacks and incidents, and ways to measure harm.

The Methodology should clarify/be sensitive to:

- interdependencies of sectors and contexts of harm (e.g., food & food security),
- link different harm categories to demonstrate cascading or linkages between types of harm,
- remain sensitive to different contexts such as target, attacker, kind of attack (for example, data theft), kind of information accessed, etc.
- how it approaches indirect harms, (e.g., taking them on a case-by-case basis). Ultimately a user should be able to understand what is meant by indirect harm,
- what it intends to measure, e.g., the actual harm or and the likelihood of harm.

On indicators, it was perceived as better to have more well-constructed indicators than fewer. There was positivity around the idea of using primary and secondary effects to prioritize the indicators that are not too complex to use. It was noted that it would be important to include "unknown values", where there is missing data.

It was mooted that an index of harms of cyberattacks and incidents would be useful, though it could be difficult to have a single value assigned to a type of cyberattack due to the various levels of impact an attack can have.

On visualizing harm indicators, it was suggested to use a spider format to comprehensively facilitate a multidimensional comparison (e.g., kinetic and cyber impact on similar scales).

Metrics are to measure harm not intent. Discussions focused on the feasibility or not of creating an indicator that deals with targeting, in particular, how broad or specific it was with respect to the initial objective.

The Institute needs to be clear on what we cannot measure.

- In its data collection methodology, the Institute will assess how to manage databases of data collection in relation to effects of cyberattacks, and ways to measure harm.
- This next phase will also focus on the primary challenges in locating applicable and usable data for each of the categories of harm, and an assessment

of the reliability of the collected data, including rating and trustworthiness.

- The Institute will also assess how the categories of harm vary across individual, organizational, international and community levels, and how indicators can be quantified or qualified in a consistent and comparable manner (and if comparability is viable).
- The Harms Methodology will clarify/ be sensitive to:
  - □ interdependencies of sectors and contexts of harm,
  - □ how to link different harm categories,
  - □ different contexts such as target, attacker, kind of attack, kind of information accessed, etc.
  - □ how it approaches indirect harms,
  - □ what it intends to measure, e.g., the actual harm or and the likelihood of harm.
  - □ the development of a potential index of harms of cyberattacks.

On a Lexicon, the experts advised that the:

- Lexicon should involve harmonizing concepts and definitions (e.g categories of harm and definition of cyber violence should list the same kinds of harm);
- definitions should be explicit and accessible for non-native English speakers and a range of stakeholders, including laypeople and policy makers.
- engaging and educating policy makers to build a framework that people can point to would be useful. The framework should not be overly scientific or complicated in nature.
- having the big picture is useful. Some cyberattacks and incidents have the same methods but ultimately have different levels of impact. That is important for policy makers to understand.

The Lexicon should carry a disclaimer, e.g. with reference to any relationship or lack thereof to legal and other technical terms that may overlap, (e.g " disclaimer: this is not a legal definition"). For example: with respect to victim, reference pre-existing definitions, and note how the Methodology definitions fit in with them.

The Institute has started to work on a Lexicon based on the advice of experts outlined above. This will be the subject of consultations over the coming months.

Definitions and Terms in Laws and Norms

Terms in norms and methodologies related to harm

There are many different terms used - often interchangeably - to explain the resulting consequences of a cyberattack or incident including “results”, “effects”, “impact”, “outcome”, “damage”, “implications”, “impairs”, and “harm”.

- Report of the United Nations Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security, (GGE) 2021, pursuant to paragraph 3 of General Assembly resolution 73/266 refers to impact, harm, consequences, implications, damages, effects[52];

- Efforts to regulate the ICT domain have led to the Framework for Responsible State Behaviour of what Member States should and should not do in the ICT environment from the perspective of international security. The Framework is the result of two decades of negotiations in various fora at the United Nations, particularly, the report of the 2021 UN Open-ended Working Group (OEWG) on developments in the field of ICTs in the context of international security and the consensus reports of the GGEs. Member States developed 11 voluntary, non-binding norms of responsible State behaviour, recommended specific confidence-building, capacity-building and cooperation measures, and that international law, in particular the Charter of the United Nations, is applicable and essential to maintaining peace, security and stability in the ICT environment. These norms refer to harmful, consequences, damages, impairs, harm.[53]

- The UNIDIR Taxonomy of Malicious Cyber Incidents references Agrafiotis et al.[54] to set an initial framework to assess harm in national contexts. Harm from cyber is understood as *“the damaging consequences resulting from cyber-events, which can originate from malicious, accidental or natural phenomena, manifesting itself within or outside of the Internet.”* This taxonomy also focuses on the primary effect on a target of a disruptive incident, as distinct from harm.

- The 2009 Report of the ICRC Expert Meeting on the Potential Human Cost of Cyber Operations[55], focused on the risk that cyber operations might cause death, injury or physical damage, affect the delivery of essential services to the population, or affect the reliability of internet services. The Report noted that *“Cyber warfare is the subject of growing concern, and there is no consensus around the question of how IHL will protect civilians against its effects”*[56]. Discussions *“helped to put the spotlight on four areas of particular concern in terms of the potential human cost of cyber operations: a) the specific vulnerabilities of certain types of infrastructure, b) the risk of overreaction due to potential misunderstanding of the intended purpose of hostile cyber operations, c) the unique manner in which cyber tools*

may proliferate, d) the obstacles that the difficulty of attributing cyber attacks creates for ensuring compliance with international law." Specific vulnerabilities of certain types of infrastructure - health care, industrial control systems i.e., electrical networks, and systemic effects to the core internet services - were highlighted[57]. On incidental civilian harm expected to be caused by a cyber operation, the view of one expert in the Report is noted in particular: *"Commanders are increasingly used to the degree of scientific sophistication reached by collateral damage estimate methodologies currently used by militaries. However, such similar methods do not exist yet for cyber operations. Another expert held that not all incidental effects would necessarily amount to legally relevant incidental civilian harm. It was also noted that the acceptable level of incidental harm and the required precautions to be taken to avoid such harm could differ depending on the actor of the type of conflict. Commanders might be more inclined to accept incidental civilian harm in armed conflicts waged for national survival than in less intense hostilities."*[58]

• The Oxford University research paper "Cyber Harm: Concepts, Taxonomy and Measurement," states that cyber harms are *"generally understood as the damaging consequences resulting from cyber-events, which can originate from malicious, accidental or natural phenomena, manifesting itself within or outside of the Internet."* The paper also uses the terms consequences, impacts and effects interchangeably.[59]

• In addition, it is important to note: the Paris Call for Trust and Security in Cyberspace is a high-level declaration that was launched in 2018 to promote stability and security in cyberspace. It is a multi-stakeholder initiative that aims to address various challenges related to the use of information and communication technologies. The Paris Call acknowledges that malicious activities in cyberspace can cause significant harm to individuals, organizations, and societies. It emphasizes the importance of protecting individuals and infrastructure from the consequences of cyber threats, such as unauthorized access, cyber espionage, and the spread of malicious software. In the context of the Paris Call, harm in cyberspace is generally understood as the adverse effects resulting from cyber incidents, including but not limited to:

- Economic harm: Cyberattacks and incidents can lead to financial losses for individuals, businesses, and governments. This can include theft of sensitive financial information, disruption of economic activities, and the costs associated with restoring systems and data.
- Privacy violations: Unauthorized access to personal data and the compromise of privacy is considered harmful.
- National security threats: Cyber threats can pose risks to the national security of countries. This may include cyberattacks targeting critical infrastructure, government systems, or military networks.

- Disruption of services: Cyberattacks can disrupt essential services, such as healthcare, transportation, and communication systems. Such disruptions can have wide-ranging consequences for societies.
- Human rights violations: The Paris Call recognizes that malicious activities in cyberspace can infringe upon human rights. This includes activities that restrict freedom of expression, interfere with democratic processes, or target individuals and groups based on their beliefs or affiliations.

Definition and measurement of harm in healthcare and environment

In healthcare, harm is a term that is immediately intuitive, implying damage and adverse consequences. To define what constitutes harm, the World Health Organisation includes the following: *"a temporary or permanent impairment, suffering, disability or loss in function or structure, which can be physical, emotional, financial or psychological, and also includes death"* (WHO 2009)[60].

In the field related to the environment, the Organization for Economic Cooperation and Development Assistance Committee (OECD-DAC) has provided a broad definition of impact as a "*positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended.*"[61]

Patient safety is also an important concept within the healthcare field, since it plays a crucial role in preventing harm. The Institute of Medicine defined patient safety as *"freedom from accidental injury"*; and the WHO as the *"absence of preventable harm to a patient and reduction of risk of unnecessary harm associated with health care to an acceptable minimum."* The concept of patient safety allows to examine the impact of incidents linked to the delivery of healthcare. A range of analyses and strategies have been implemented by healthcare practitioners in order to mitigate the effects of incidents. The implementation of harm measurement approaches has also been introduced as a means to achieve this objective. In this context it is instructive to highlight terms and definitions for grading patient safety incidents:

- *None/Insignificant:* impact prevented (incident that had the potential to cause harm but was prevented, resulting in no harm); impact not prevented (incident that ran to completion but no harm occurred to people)
- *Low/Minor:* incident that required extra observation or minor treatment and caused minimal harm, to one or more persons
- *Moderate:* incident that resulted in a moderate increase in treatment and which caused significant but not permanent harm, to one or more persons
- *Severe/Major:* incident that appears to have resulted in permanent harm to one or more persons

- *Catastrophic/ Death:* incident that directly resulted in the death of one or more persons.

WHO also provides the Surveillance System for attacks on Health Care (SSA), a mechanism for monitoring "kinetic" attacks against healthcare facilities. The SSA, is a web-based reporting system and platform, that comprises standardized data describing an attack, the type of attack, and how health resources are affected including health facility, transport, personnel, patients, supplies, warehouses. The impact on people is measured according to two metrics: injuries and death, and quantified by the total number of victims injured or dead by sex, age group, and type.

The SSA classifies an attack as direct or indirect, targeted or untargeted. Any report that is considered to be either direct or targeted in character is immediately labeled as an attack on health care, with WHO reviewing indirect and non-targeted occurrences on a case-by-case basis.

Notions of harm in insurance and domestic laws

These areas present a range of three different notions of harm:

A Comprehensive List of Harms - This concept of harm is implicitly defined in terms of specific inherently harmful actions. These quantifications of harm rely on comprehensive lists of possible harmful scenarios, and an associated measure of severity for each scenario. For example, an insurance policy might identify loss of vision as a specific scenario, as well an identified severity of the harm quantified by the degree of cover available for that injury. Other examples include UK criminal law, which identifies certain acts, e.g coercion, with severity of harm expressed in terms of specific associated penalties like fines or jail-time. This conception of harm implies a range of different harms, including physical, psychological and deprivational, across a range of scenarios.

Harm as Costs Incurred - This approach defines harms in terms of the services that are required after the harm. Psychiatric harms require psychiatric treatment, physical harms require medical treatment, and reputational harms require brand management. Thus these could be considered three different kinds of harm, and the degree of harm is determined by the cost of the required service. Examples include insurance for psychiatric, reputational, and intellectual property insurance policies.

Reasonable Definitions of Harm - This approach sets some broad definitions of harms in context, and leaves specific determinations to be made on a case by case basis, based on reasonable understandings of the terms. For example, in the context of forced labour, the United States Code defines harm as "*physical or nonphysical... harm, that is sufficiently serious...to compel a reasonable person of the same background and in the same circumstances to perform ...labor...to avoid incurring that harm.*"[62]

These approaches present a range of coarse and fine-grained conceptions of harm, and demonstrate an array of viable approaches, including to non-physical harm, that are well tested and workable. They also demonstrate that there are existing reliable mechanisms for assessing harms beyond financial and physical, relying on specialist understandings of non-physical harms, such IP insurers and psychiatrists, and general shared understanding of common terms.

Cybercrime law and notion of harms

Concerning cyber-specific law, the notion of harm is largely absent from the Council of Europe Convention on Cybercrime (also known as the Budapest Convention) and other regional conventions which deal specifically with cybercrime. The most prominent after the Budapest Convention is the Malabo Convention (Convention on Cyber Security and Personal Data Protection) of the African Union which indeed does not mention harm or impact.

From the Budapest Convention: Article 4 – Data interference:

> 1. Each Party shall adopt such legislative and other measures as may be necessary to establish as criminal offences under its domestic law, when committed intentionally, the damaging, deletion, deterioration, alteration or suppression of computer data without right.
>
> 2. A Party may reserve the right to require that the conduct described in paragraph 1 result in serious harm.

From the 2nd Additional Protocol of the Budapest Convention: Data security and security incidents:

> a. Each Party shall ensure that it has in place appropriate technological, physical and organisational measures for the protection of personal data, in particular against loss or accidental or unauthorised access, disclosure, alteration or destruction ("security incident").
>
> b. Upon discovery of a security incident in which there is a significant risk of physical or non-physical harm to individuals or to the other Party, the receiving Party shall promptly assess the likelihood and scale thereof and shall promptly take appropriate action to mitigate such harm.

Computer-related offenses that cause personal harm, such as cyberharassment, cyberbullying and cyberstalking are part of certain domestic law. Some countries, without expressly mentioning harm, criminalize the use of a computer to send any data that intends to cause, or is reckless as to whether the sending of the data causes, *annoyance, inconvenience, distress, or anxiety,* to that person or any other person. Therefore, particular attention is paid to the consequences of the offense so that it can be constituted.

The Draft UN Cybercrime Convention deals with non-consensual dissemination of intimate images (Article 15), para. 5: State Party may require the intent to cause harm before criminal liability attaches.

The notion of harm has been part of the discussions taking place at the

UN Ad Hoc Committee to Elaborate a Comprehensive International Convention on Countering the Use of Information and Communications Technologies for Criminal Purposes (also known as the UN Cybercrime Convention). Guiding questions for State interventions in regard to the provisions on criminalization sought views on whether the proposed conducts must result or be intended to result in a specific or serious harm, or material damage, in order to be considered as an offense.

The States were also requested to share their views on how harm should be defined.[63] The revised draft text of the UN Cybercrime Convention (A/AC.291/22/Rev.1) considers harm as a threshold for criminalization in two instances.

States may require that interference with computer data such as its damaging, deletion, deterioration, alteration or suppression results in serious harm to be considered a criminal offense (Art. 8). The intent to cause harm may also be required before criminal liability attaches in offenses related to non-consensual dissemination of intimate images (Art. 15)[64].

The UN Cybercrime Convention includes provisions relevant to the protection of witnesses who give testimony or provide information concerning criminal offenses (Art. 33) and assistance to and protection of victims of cybercrime (Art. 34). States are requested to take appropriate measures to protect witnesses and victims from potential retaliation or intimidation, including procedures for the physical protection of witnesses. States are further requested to establish procedures to provide access to compensation and restitution for cybercrime victims and measures to provide assistance to victims, including for their physical and psychological recovery while taking into account the particular circumstances and needs of victims. However, the standards for the protection of witnesses and victims of cybercrime are subject to domestic law and the treaty does not oblige States to meet international human rights standards.

Notion of harm and victim in international law

In international law, rules exist on the responsibility of States for internationally wrongful acts, which apply to acts committed in cyberspace. States also have a responsibility to prevent certain acts committed by non-State actors within their jurisdiction, which may include cyber crime. These rules provide a basis for international responsibility in case human harm is unlawfully caused in cyberspace.[65]

The **notion of harm** barely exists in public international law. Indeed, the conditions for engaging the international responsibility of a State do not take into account the existence of harm. The responsibility of a State for an internationally wrongful act is triggered if the breach of an international obligation is attributable to the State. However, demonstrating the existence of a causal relationship between the harm suffered and the act giving rise to liability is a condition of the obligation to make reparation under public international law.

This absence is explained by the fact that, traditionally, international law and its interstate structure was created to respond to State's interests and goals and did not pay attention to victims nor to harm. Individuals were only taken into account in some particular fields of international law. On the one hand, in international human rights law, there is a victim when the State is the author of the violation of international obligations, the individual thus claims a violation against the State. On the other hand, in international criminal law and IHL, victims are acknowledged when other individuals are the author of the breach.

A body of law in which the concept of harm, by contrast, is very present is international environmental law. In this branch of law, the no-harm rule is a widely recognised principle of customary international law whereby a State is duty-bound to prevent, reduce and control the risk of environmental harm to other states[66].

The **notion of victim** remains very specific to each body of law, which has its own definition and its own regime. Therefore, different categories of victims could be defined.

A first category of victims would be the victims of crime and abuse of power. In 1985, the UN General Assembly adopted the Declaration of Basic Principles of Justice For Victims of Crime and Abuse of Power.[67] This declaration defined victims as *"persons who, individually or collectively, have suffered harm, including physical or mental injury, emotional suffering, economic loss or substantial impairment of their fundamental rights, through acts or omissions that are in violation of criminal laws operative within Member States, including those laws proscribing criminal abuse of power"* and recognizes that victims have rights and needs. The declaration provides for example, for access to justice and fair treatment, restitution, compensation, and assistance.

Another category is for the victims of violations of human rights law. The Basic Principles and Guidelines on the Right to a Remedy and Reparation for Victims of Gross Violations of International Human Rights Law and Serious Violations of

International Humanitarian Law[68], adopted by the UN General Assembly in 2005 has led to changes in victim's right to compensation. According to these principles, compensation must be paid for the damages such as: physical or mental harm, moral damage, or lost opportunities, including employment, education and social benefits. Human Rights Law contains specific provisions regarding the protection of victims of crime such as the right to a fair trial or the prohibition of discrimination. The jurisprudence of human rights jurisdictions has specified that States have a positive obligation to ensure the enjoyment of the victim's fundamental rights.

Victims of violations of International Criminal Law would be another category. The adoption of the Rome Statute is characterized by a victim-centered approach. Victims can thus be defined as *"those who have suffered harm as a result of the commission of any crime within the jurisdiction of the Court. Victims may include individual people, but also organizations or institutions that have sustained direct harm to any of their property which is dedicated to religion, education, art or science or charitable purposes, and to their historic monuments, hospitals and other places and objects for humanitarian purposes"*[69]. In particular, the Rome Statute entitles victims the right to participation in trials and the right to reparations.

UN Charter

The notions of use of force, armed attacks and aggression are found in the UN Charter but does not specifically refer to cyber means or operations. States have been asked to clarify their positions with regard to the restrictions and limits imposed on the use of operations.

- **Use of force**

According to States, there is no doubt that the prohibition of the use of force arising from Article 2(4) of the UN Charter applies to cyber operations. They consider that for a cyber operation to amount to a prohibited use of force, its scale and effects must be comparable to those caused by the use of conventional weapons. They generally agree that it is so when the operation causes death, physical harm or injury to persons or substantial material damage to the victim state's objects and/or state functioning.

Some States[70] rely on a list of non-exhaustive factors to assess if a cyber operation reaches the level of use of force. These factors are severity, immediacy, directness, invasiveness, measurability of effects, military character, State involvement, presumptive legality.

States[71] have given examples of cyber operations that would be considered as a use of force, which may include:

- An action in cyberspace that leads to: a permanent and significant damage of a power plant, a missile defence system deactivation or taking control over an

aircraft or a passenger ship and causing an accident with significant effects,[72]

- A cyber operation permanently disabling operating systems controlling critical infrastructure, such as an electrical grid or a water and sanitation station,[73]
- Penetrating military systems in order to compromise States defence capabilities, or financing or even training individuals to carry out cyberattacks against the State,[74]
- A major offensive cyber operation, destroying servers used by the State's military headquarters, resulting in the State's inability to communicate with naval vessels operating in international waters off the coast of a foreign State, and will take several months to replace the destroyed servers, at substantial cost,[75]
- A cyber operation causing severe disruption to the functioning of the State such as the use of crypto viruses or other forms of digital sabotage against governmental or private power grid- or telecommunications infrastructure, or cyber operations leading to the destruction of stockpiles of COVID-19 vaccines,[76]
- Hacking into the computers of the railroad network of another State and programming the controls in a manner that is expected to cause a collision between trains.[77]

However, some points are still controversial or under debate, for example, It is not yet clear how long-term and/or indirect impact is taken into account,[78] or whether the economic effects of a cyber operation should be included in the definition of use of force. In any event, the economic effects must be significant, such as the collapse of a State financial system, or parts of its economy.[79] Another area requiring clarification is the qualification of unlawful use of force when there is no physical damage resulting from a cyber operation but only a loss of functionality.[80]

- **Armed attack and aggression**

The term 'armed attack' is used in Article 51 of the UN Charter. According to this article, a State may use self-defence when it is the object of an armed attack. This is an exception that allows a State to use force.

As noted by the International Court of Justice (ICJ), armed attacks are the 'most grave' forms of use of force[81] and "scale and effects" are to be considered when determining whether particular actions amount to an 'armed attack'.[82] In this way, the qualification of an armed attack does not depend on the means but on the consequences of the operation.

States generally therefore consider that a cyber operation reaches the threshold of an armed attack if its scale and effects are comparable to those of a kinetic attack. Thus, when a cyber operation causes substantial death or injury of people or considerable material damage or destruction of property, the victim State enjoys the right of self-defence.

States reaffirm that this right to use force, including through cyber means, should be exercised in line with international

law, namely the principle of necessity and proportionality.[83] The right of self-defence can be exercised through either cyber or conventional means.

The elements on which there is still no general consensus are:

- The qualification of a cyberattack as an armed attack if it does not cause fatalities, physical damage or destruction yet nevertheless has very serious non-material consequences.[84]
- Cyberattacks which do not reach the threshold of an armed attack when taken in isolation could be categorised as such if the accumulation of their effects reaches a sufficient level of gravity.[85]
- The applicability of the right of self-defence to an armed attack perpetrated by a non-state actor whose action is not attributable to a State.[86]
- The right to use self-defence before the attack actually occurs.[87]

Examples of a cyber operation amounting to an armed attack, triggering the right of self-defence could be:

- A cyberattack leading to the disabling of an air traffic control system which causes planes to crash or an interference with the operating system of a power station, which causes serious physical damage,[88]
- An operation in cyberspace that caused a failure of critical infrastructure with significant consequences or consequences liable to paralyse whole swathes of the country's activity, trigger technological or ecological disasters and claim numerous victims,[89]
- A cyber activity that disables the cooling process in a nuclear reactor, resulting in serious damage and loss of life.[90]

- **Sovereignty**

Sovereignty is a fundamental principle of international law. The internal aspect of sovereignty implies that States enjoy the exclusive jurisdiction over all persons, property and events within their territory, within the limits of their obligations under international law.[91] Any interference by another State in inherently governmental functions is prohibited. On the other hand, the external aspect pertains to the international equal rights and duties of a State in its relations to other States.[92]

It is widely recognized that the principle of sovereignty applies to activities in cyberspace. States consider that the violation of the principle of sovereignty by a cyber operation is a non-consensual intrusion in the computer networks and systems of another State. The position of each State reveals a consensus that to amount to a violation of the principle of sovereignty, the effects of a malicious cyber operation must be:

- Either significant intentional or unintentional harmful effect on cyber infrastructure components or on persons or other infrastructures (whether public or private). Such effects could be a loss of functionality, physical damage, or modification or deletion of information, that necessitate the repair or replacement of physical components

of cyber infrastructure. But a cyber operation that only requires a rebooting or reinstallation of an operating system is likely not a violation of sovereignty. Besides, States generally agree that all intrusions are not a violation of sovereignty, espionage by cyber means should remain authorized for instance.

- Or the interference with data and services that are necessary for the exercise of inherently governmental functions (irrespective of any physical or non-physical effects). These are health care services, law enforcement agencies, administration of elections, tax collection, national defence and the conduct of international relations, foreign policy, critical infrastructure or company of public interest, energy, water, and sanitation facilities.

Some States have given examples of cyber operations that could amount to a violation of sovereignty[93]:

- Cyber operation preventing the proper functioning of ICT networks, services or systems of public entities, or a theft, erasure or public disclosure of data belonging to such entities[94],
- Cyber operation against warships, ships owned by a State and used only for government or non-commercial service, or against State aircraft, [95]
- Interfering with a State's democratic processes, such as elections, responses to a national security or health emergency, such as the COVID-19 pandemic, and its choice of foreign policy,[96]
- A cyber activity that interrupts health care delivery by blocking access to patient health records or emergency room services, resulting in risk to the health or life of patients,[97]
- A cyber operation against an industrial control system at a petrochemical plant that led to a malfunction and a subsequent fire.[98]

International Humanitarian Law (IHL) and harm

IHL and its principles of distinction, proportionality, precaution, military necessity and humanity restrict the use of cyber means and methods used in armed conflict.

- **Notions of harm, violence and attack**

In IHL, the notion of harm is related to the notion of foreseeability. When preparing or launching the attack there must always be an assessment ex ante of the potential harm it could cause according to the information available at the time. To assess the final harm from the attack, one cannot only refer to the IHL notion of foreseeable harm and damage: it is important to assess the consequences of the attack on the civilian population, whether or not those operations were lawful under IHL.

Civilians and civilian objects are protected under IHL unless they turn into military objectives (Article 52(2) of Additional Protocol I for civilian objects, Article 51(3) for civilians). Indeed, civilians are protected "unless and for such time as they take a direct part in hostilities." According to the ICRC 2009 Interpretative Guidance on Direct Participation in Hostilities, *"Persons take a direct part in hostilities when they commit acts aimed at supporting one party to the conflict by directly causing harm to another party, either by directly inflicting death, injury or destruction, or by directly harming the enemy's military operations or capabilities."* Three criteria are thus required: the belligerent nexus, a minimum threshold of harm and a direct causation between the act and the harm.

The notion of attack is defined in Article 49(1) of AP I as *"acts of violence against the adversary, whether in offence or in defence."* In its commentary of the article, the ICRC describes an attack as *"the use of armed force to carry out a military operation at the beginning or during the course of armed conflict."* The second paragraph of Art. 49 refers to any *"land, air or sea warfare that may affect the civilian population, individual civilians or civilian objects on land."*

It is widely accepted that the notion of violence in the definition of attacks can refer to either the means of warfare or their effects, meaning that an operation generating violent effects can qualify as an attack even if the means used to bring about those effects are not violent as such. It is also widely accepted that cyber operations expected to cause death, injury or physical damage constitute attacks under IHL. For a number of States, as well as the ICRC, during an armed conflict an operation designed to disable a computer or a computer network constitutes an attack under IHL, whether the object is disabled through kinetic or cyber means.

Although non-binding, the Tallinn Manual on the International Law Applicable to Cyber Operations Rule 30 defines a cyberattack as a *"cyber operation, whether offensive or defensive, that is reasonably expected to cause injury or death to persons or damage or destruction to objects."* This definition focuses thus on the consequences -

the harm - of the attack rather than the way it is conducted. The notion of consequential damage, destruction, injury or death.

Therefore, to be considered as an attack in the meaning of the laws of war, its consequences must reach a certain threshold of harm, which is not clearly defined, although excluding *"de minimis damage or destruction"*. Clarity would need to be provided on what de minimis means. The Tallinn Manual 2.0 broadens the definition regarding the target, which does not need to be the adversary, to make sure that all civilians are included, but also regarding the effects. Indeed, they also encompass *"serious illness and severe mental suffering that are tantamount to injury."*

There is debate about the notion of loss of functionality of an object i.e. to make it dysfunctional without physically damaging it, which is feasible in cyberspace. "*The most permissive approach is to consider that cyber attacks are only those operations that cause violence to people or physical damage to objects. A second approach is to make the analysis dependent on the action necessary to restore the functionality of the object, network or system. A third approach is to focus on the effects that the operation has on the functionality of the object.*" [99]

- **Conduct of hostilities**

If the operation amounts to an attack, the commander preparing and launching it is bound by the rules on Conduct of Hostilities enshrined in the 1907 Hague Regulations, include the Additional Protocol I of 1977 to the Geneva Conventions, and customary international humanitarian law. They detail the three main sets of rules on distinction, proportionality, and precautions which aim to protect the civilian population.

IHL provisions also related to the means and methods of warfare. Article 35 of Protocol I states for example that *"it is prohibited to employ weapons, projectiles and material and methods of warfare of a nature to cause superfluous injury or unnecessary suffering"* and also prohibits *"widespread, long-term and severe damage to the natural environment"*. These provisions clearly demonstrate the will to prevent and minimize as much as possible the harm caused to people as a consequence of an armed conflict.

Additional Protocol I also includes a provision on new weapons (which can therefore apply to cyber weapons), stating that the party is under the obligation to determine if the employment of this new weapon will be prohibited by the Protocol or other international law.

IHL also sets out fundamental principles for the conduct of hostilities in order to limit the consequences of war on individuals. Thus, attacks must be proportionate and must respect the principle of distinction[100] between civilians and military personnel (Art. 48 Protocol I), the principle of precaution (Art. 57 Protocol I) and the principle of necessity.

- The principle of proportionality is defined in Article 51(5)(b) of AP I: an attack is

prohibited if it *"may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated."*

If the notion of excessiveness is not clearly defined, this principle binds the commander to assess, ex ante and based on the information available at the time, what incidental civilian harm is expected to be caused, in relation to the concrete and direct military advantage gained.

It is transcribed in Rule 113 of the Tallinn Manual 2.0. In order to assess proportionality, the commander must take into consideration not only the direct harm resulting from the attack, but also the indirect effects which comprise *"the delayed and/or displaced second-, third-, or higher-order consequences of action, created through intermediate events or mechanisms."*

- Under the rules of precautions, attackers must adhere to Article 57 of AP I, taking all feasible precautions during planning and execution. This involves verifying the target's legitimacy and minimizing civilian harm, with a mandate to cancel or suspend an attack if it violates distinction or proportionality principles. Effective warnings to civilians are required unless circumstances prevent it. While most precautionary rules apply to attackers, defenders also have obligations, such as passive precautions under Article 58, which may involve limiting interconnectivity or developing shields in cyberspace to protect civilian systems.

The ICRC Expert Meeting on The Potential Human Cost of Cyber Operations, focused in particular on the risk that cyber operations might cause death, injury or physical damage, affect the delivery of essential services to the population, or affect the reliability of internet services. It looked at the specific characteristics of cyber tools, how cyber threats have evolved, and the cyber security landscape. The Report highlighted specific vulnerabilities of certain types of infrastructure: cyberattacks that may affect the delivery of healthcare, industrial control systems, or the reliability or availability of core internet services. Apart from causing substantial economic loss, the Report states that, cyber operations can harm infrastructure in at least two ways. First, they can affect the delivery of essential services to civilians, as has been shown with cyberattacks against electrical grids and the healthcare sector. Second, they can cause physical damage. Health care infrastructure is particularly vulnerable, with potentially serious consequences for health and life.

The Report noted that while the risk of human cost based on current observations does not appear extremely high, especially considering the destruction and suffering that conflicts always cause, the evolution of cyber operations still merits close attention due to existing uncertainties and the rapid pace of change.[101]

- Attack and violence

The notion of attack is defined in Article 49(1) of Additional Protocol I to the Geneva Conventions, as *"acts of violence against the adversary, whether in offence or in defence."* In its commentary of the article, the ICRC describes an attack as *"the use of armed force to carry out a military operation at the beginning or during the course of armed conflict."* The second paragraph of Article 49 refers to any *"land, air or sea warfare that may affect the civilian population, individual civilians or civilian objects on land."*

The act of violence and the notion of physical force to carry out a military operation are concepts that do not translate easily to attacks in cyberspace. Cyberspace knows no territory beyond the ones from which the attack is launched or where it has effects. However, a cyberattack may affect persons and objects on land, meeting the requirements of the provision.

It is widely accepted that the notion of violence in the definition of attacks can refer to either the means of warfare or their effects, meaning that an operation generating violent effects can qualify as an attack even if the means used to bring about those effects are not violent as such. It is also widely accepted that cyber operations expected to cause death, injury or physical damage constitute attacks under IHL. For a number of States, as well as the ICRC, during an armed conflict an operation designed to disable a computer or a computer network constitutes an attack under IHL, whether the object is disabled through kinetic or cyber means.

At the international level, even though it is agreed that IHL applies to cyberspace and restricts the use of cyber capabilities as a means and method of warfare during an armed conflict, there is a need for clarity on the limits that IHL imposes on the use of cyber operations due to the complexity of this realm and the challenges in terms of applicability, and accountability. Clarifications related to the interpretation of the rules are still required by States, and intergovernmental discussions have been taking place in the United Nations mandated process working on the development of the regulations in cyberspace, currently through the ongoing OEWG on security of and in the use of information and communications technologies (2021-2025). A small number of States have done so, some of the positions have been included herewith.

In war time, another convention can be applied that refers to harm to people. It is the **UN Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons** which may be deemed to be Excessively Injurious or to have Indiscriminate Effects (of 1980). The aim is to set rules on the use of weapons in order to minimize the harm caused by these weapons to individuals. The provisions are complementary to Article 35 of Protocol I to the Geneva Convention and therefore prohibits weapons that injure by fragments which in the human body escape detection by X-rays, and the use of booby-traps. The use of mines or incendiary weapons are also restricted.

Rome Statute of the International Criminal Court on the notion of harm

The notion of harm appears in the Rome Statute of the International Criminal Court (ICC) as an element of the crime of genocide.

> Article 6: *"For the purpose of this Statute, "genocide" means any of the following acts committed with intent to destroy, in whole or in part, a national, ethnical, racial or religious group, as such:*
> *(a) Killing members of the group;*
> *(b) Causing serious bodily or mental harm to members of the group;*
> *(c) Deliberately inflicting on the group conditions of life calculated to bring about its physical destruction in whole or in part;*
> *(d) Imposing measures intended to prevent births within the group;*
> *(e) Forcibly transferring children of the group to another group."*

Note: Only acts causing "serious" bodily or mental harm are acts of genocide. Serious bodily harm is determined on a case-by-case basis and it has been decided that it is *"harm that seriously injures the health, causes disfigurement or causes any serious injury to the external, internal organs or senses"*[102]. Other acts found to result in bodily harm include inhumane and degrading treatment[103], deportation, enslavement, starvation, persecution, and interrogations combined with beatings.

There is far less jurisprudence to define mental harm. Serious mental harm must involve *"some type of impairment of mental faculties or harm that causes serious injury to the mental state of the victim"* [105]. Courts recognised as mental harm the threat of death and knowledge of impending death, acts causing intense fear or terror, surviving killing operations, forcible displacement, and mental torture[105].

The notion of harm can also be found in Article 8 of the Rome Statute criminalizing war crimes and violations of International Humanitarian law. As an example, *"Intentionally launching an attack in the knowledge that such attack will cause incidental loss of life or injury to civilians or damage to civilian objects or widespread, long-term and severe damage to the natural environment which would be clearly excessive in relation to the concrete and direct overall military advantage anticipated"* (art. 8 (2)(b)(iv)) is considered a war crime.

Furthermore, in August 2023, the ICC's chief prosecutor declared the potential prosecution of cybercrime before the Court where the case is sufficiently grave. He explained that conduct in cyberspace may potentially amount to war crimes, crimes against humanity, genocide, and/or the crime of aggression. He also added, *"The digital front lines can give rise to damage and suffering comparable to what the founders of the ICC sought to prevent. Cyber warfare does not play out in the abstract.*

Rather, it can have a profound impact on people's lives. Attempts to impact critical infrastructure such as medical facilities or control systems for power generation may result in immediate consequences for many, particularly the most vulnerable. Consequently, as part of its investigations, my Office will collect and review evidence of such conduct"[106].

Concluding Remarks and Next Steps

The Expert Meeting was an important first milestone for the CyberPeace Institute in publicly sharing its work to develop a standard data driven Harms Methodology and metrics to understand, track, and measure the harm from cyberattacks and incidents. The feedback received from experts during this meeting was insightful and will guide the Institute as it continues the development of this Harms Methodology.

The meeting allowed the CyberPeace Institute to confirm much of its research which was submitted to the Experts, to nuance some of its thinking benefitting from the feedback received, and to move forward and confirm next steps.

Our research and feedback from the Experts Meeting confirmed the complexity of the development of such a methodology and metrics due to the broad range of considerations that need to be factored into this work. However, the important contribution that such a Methodology could make to understanding and measuring harm more comprehensively was also underlined.

The publication of this Report of the Experts Meeting will enable the Institute to engage with experts who could not attend this meeting, and to broaden our outreach to a range of additional stakeholders for their insights. The Institute will leverage this Report to engage with States and civil society actors over the next months.

In parallel, the Institute will continue its research focusing particularly on an ontology of terms for data collection, and operationalising the definition through continued work on indicators and metrics, including through assessments of further case studies of a range of types of cyberattacks and incidents. Case studies continue to enable the Institute to explore indicators and metrics, and to test data collection needs.

In this regard, the Institute is working on a pilot project and AI modeling based on known features of the harm caused by a cyberattack - together with other details such as claims by perpetrators or threat actors. This modeling entails leveraging AI as a diagnostic tool that then gives possibilities of type of attack, speed of spread, the “knock-on” human impact, origin, type of attack, intent, etc. The focus will be on instructing the tool to undertake the analysis and write the outcome in the format given by the definition of the Theory of Violence. The findings of this research will be consolidated into background documents for a further consultation meeting with experts.

Meanwhile, any feedback on this Report and ongoing work can be shared with the Report authors at clindsey@cyberpeaceinstitute.org and kamdouni@cyberpeaceinstitute.org. With thanks also for the contributions of Gwyn Glasser, Solène Poleart and Pavlina Pavlova.

The Institute welcomes engagement and collaboration on this work.

About the Cyber Peace Watch

The CyberPeace Watch (Watch) is an interactive, online platform being developed by the CyberPeace Institute, providing an easily accessible baseline of data to understand and share knowledge about cyberattacks, including the analysis of threats, harms, and related paths for accountability. The platform's goal is to assess cyber peace based on evidence of the harms caused by cyberattacks and the actions taken by States and other relevant actors to strengthen responsible behavior in cyberspace.

The beta versions of this new Platform are the Cyber Incident Tracers #Health published in 2021 and the Cyber Attacks in Times of Conflict Platform #UKRAINE (CATC) published in 2022. The latter is a platform on attacks on critical infrastructure sectors essential to civilians, on threat actors, and provides an overview on law and policy. The data collection is currently being expanded to monitoring other attacks carried out by threat actors without a specific geographic scope.

The Cyber Incident Tracers are online platforms with accessible baselines of data on cyberattacks, created to consolidate evidence-based insights that demonstrate the full complexity, scale, and impact that cyberattacks are having on people. Every research project begins with the setting of clear intelligence requirements and the definition of research questions that need to be answered. This ensures our research stays within scope, respects ethical research principles and avoids mission creep.

Manually, automatically or a combination of both methods is used to collect data on cyberattacks from primary data sources, open sources and closed sources. When combined together, this collection gives us a more comprehensive understanding of the cyber risks faced by vulnerable communities. Using data pipelines we clean and normalize data and evaluate its relevance and reliability to transform it from its raw form to exploitable information usable for analysis. This step of our intelligence cycle requires close collaboration between our analysts and technical engineers.

From data discovery, statistical analysis, Social Network Analysis to geotemporal analysis, we find hidden connections within large datasets. Using data visualization and analysis tools, including dashboards and graphical link analysis software, our analysts can connect information from disparate sources to find the answers to our research questions.

Complex analysis must be accompanied by simple storytelling. Developing data-visualization platforms tailored to each research project and publishing clear reports and infographics allow us to communicate our findings and engage, including in public policy negotiations.

The CyberPeace Watch online platform, and a series of accompanying reports, will be launched in 2024.

Timeline of CyberPeace Watch Initiatives

2022-23	Research on Harms Methodology
November 2023	First Expert Meeting on Harms Methodology
December 2023	Publication of Report on Expert Meeting and follow up
	Launch bilateral consultations on sidelines of OEWG New York
	Statement to OEWG
January-March 2024	Consultations on Expert Meeting Report
	Continued Research
	Preparation of Expert Meeting II
April 2024	Expert Meeting II
	Publication of Expert Meeting Report
	Launch of CyberPeace Watch Platform
June 2024	Publish draft Harm Methodology for consultation
Autumn 2024	Publish final Harm Methodology and supporting documentation

Annexes

Annex 1 - Participants to Experts Workshop, November 2023

- Aad Imad, Ecole Polytechnique Fédérale de Lausanne, Center for Digital Trust (C4DT)
- Abed Saif, World Health Organisation
- Benincasa Eugenio, Ecole polytechnique fédérale de Zurich, Center for Security Studies (CSS)
- Boichat Gabriel, Delegation of the Catalan Government to Switzerland
- Bundt Maya, Board member, CyberPeace Institute
- Buzatu Anne-Marie, ICT4Peace
- Castella Grégoire, Ecole Polytechnique Fédérale de Lausanne, EssentialTech Center
- Dominioni Samuele, United Nations Institute for Disarmament Research
- Doumanian Lucie, SMEX
- Dr Allain Loos Sophie, World Health Organisation
- Francisco Carolyn, MITRE Corporation
- Gual Carme, Generalitat de Catalunya, General Directorate for Digital Society
- Grelin Guillaume, Représentation permanente de la France auprès de l'ONU à Genève
- Harmes Robert, United Kingdom Department for Science, Innovation & Technology
- Hernandez Elsa, United Kingdom Department for Science, Innovation and Technology
- Hudson Alexander, International IDEA
- Ito Yurie, CyberGreen Institute
- Janovsky Marek, Czech Republic Permanent Mission to the UN in Geneva
- Kastelic Andraz, United Nations Institute for Disarmament Research
- Knobel Thomas, University of Lucerne
- Kobel Vivienne, Centre for Feminist Foreign Policy
- Lokhorst Natha, The Hague Centre for Strategic Studies
- Lyons Josh, GEO incognita
- MacColl James, Royal United Services Institute
- Morgan Lawrie, United Kingdom Department for Science, Innovation, and Technology
- Philibert Martin, International Telecommunication Union
- Pytlak Allison, Stimson Center
- Robin Coupland, Independant
- Rodrigues Stacy, Luxembourg Ministry of Foreign and European Affairs, Luxembourg
- Sean Cordey, International Committee of the Red Cross
- Shires James, Chatham House, United Kingdom
- Taback Nathan, University of Toronto
- Veit Meredith, Business & Human Rights Resource Centre
- Wiedemar Sarah, Ecole polytechnique fédérale de Zurich, Center for Security Studies (CSS)
- Wille Christina, Insecurity Insight

Annex 2 - Case Studies

Case Study 1 - Viasat

About VIASAT

Viasat is an American communications company based in Carlsbad, California, with additional operations across the United States and worldwide. Viasat is a provider of high-speed satellite broadband services and secure networking systems covering military and commercial markets.

Overview

On February 24th, 2022, the day of Russia's invasion into Ukraine, a cyberattack disrupted broadband satellite internet access. This attack disabled modems that communicate with Viasat Inc's KA-SAT satellite network, which supplies internet access to tens of thousands of people in Ukraine and Europe. Researchers from SentinelLabs believe that the attack was the result of a new strain of wiper malware called "AcidRain'' that was designed to remotely erase vulnerable modems and routers.[107] [108] According to the NSA, in an effort to keep specific modems offline, hackers flooded Viasat's systems with requests, overloading their systems [109]. Viasat agreed with this assessment, and in a later statement said they believed the purpose of the attack was to interrupt service rather than to access data or systems. The United State's assessed "*...that Russia launched cyber attacks in late February against commercial satellite communications networks to disrupt Ukrainian command and control during the invasion, and those actions had spillover impacts into other European countries*." [110]

Impact

As the attack impacted telecommunications systems, it did not just have the potential to threaten government or military objects, but rather it also impacted the civilian population and civilian objects both in Ukraine and beyond when they experienced a loss of internet access and possible disruptions to systems in the energy sector. Some reported that their internet access was offline for more than two weeks.

The attack on Viasat also impacted a major German energy company who lost remote monitoring access and control to over 5,800 wind turbines across 1217 farms, and in France nearly 9,000 subscribers of a satellite internet service provider experienced an internet outage. In addition, around a third of 40,000 subscribers of another satellite internet service provider in Europe (Germany, France, Hungary, Greece, Italy, Poland) were affected. Overall, this attack impacted several thousand customers located in Ukraine and tens of thousands of other fixed broadband customers across Europe.

Attribution

A first technical attribution was conducted and publicly disclosed by SentinelLabs at the end of March 2022, as they found that AcidRain presented developmental similarities with a 2018 VPNFilter campaign previously attributed to the Russian government.[111]

Months later, on May 10, the EU and the Five Eyes governments consisting of the United States, United Kingdom, Australia, New Zealand, and Canada, released public statements attributing AcidRain to the Russian military intelligence (GRU) and linking it to multiple families of destructive wiper malware, including WhisperGate, targeted on the Ukrainian government and private sector networks. Further specific national statements aligning with this attribution were made by the ministries of foreign affairs of Estonia, Denmark, Ireland, the Netherlands, Norway, Austria, Germany, Czechia, Italy, Finland, Romania, Poland, and France. This consistent response by many governments is an important step in the practice of political attribution of cyberattacks and greatly contributes to the development of states' practice in this sense.

In addition, many of the statements presented references and allegations to Russia's violations of the normative framework for responsible state behavior in cyberspace, as established through the consensus reports of the UN Group of Governmental Experts (UNGGE) and reaffirmed by the previous Open Ended Working Group (OEWG). According to these semi-collective attributions, both the targeting of critical infrastructures and the spillover effects on civilians not being directly involved in the conflict are undermining the rules-based international order. Thus, the public statements that followed the Viasat cyberattack contribute to a certain extent to improve the understanding of states' view on how international law and the UN normative framework applies to cyberspace.

Public attribution: The European Union and its Member States, the UK, and the USA have politically attributed this attack to the Russian Federation. [112] [113] [114]

Nation state actor attributed to perpetrating the cyberattack: Russian Federation - specifically the Russian foreign military intelligence agency (GRU)

Case Study 2- Vastaamo

About Vastaamo

Vastaamo was a Helsinki-based private psychotherapy center founded in 2008 that provided private mental-health services to its patients. It was a firm with 28 therapy centers throughout Finland.Vastaamo operated as a subcontractor for Finland's public health system.

Overview

In late September 2020, the Vastaamo Psychotherapy Center was made aware that its systems were breached on two separate occasions in November 2018 and March 2019. Attackers did not contact Vastaamo until September 2020, at which point they demanded a ransom payment of 40 Bitcoins (~450,000 EUR). On October 21, after Vastaamo refused to pay the ransom, the attackers began posting batches (100 records a day) of patient records on underground forums and requesting that patients pay 500 EUR to have their information taken offline. On October 24, 2020 Vastaamo reported that patients and employees began to receive extortion emails from the threat actor(s) requesting bitcoin payments, otherwise data would be published online. Since these events occurred, over 25,000 criminal complaints have been filed by victims and a criminal investigation has concluded in Finland with suspect Julius Kivimäki on remand since February 2023.

Impact

As a result of this hack, approximately 36,000 patient records, including juveniles, were stolen. These records contained highly sensitive personal data including names, contact details, social security numbers and records of therapy sessions of some of the most vulnerable in society as well as the healthcare professionals who treated them. Around 30,000 people are believed to have received the ransom demand and over 25,000 reported it to the police. A 10-gigabyte data file containing private notes between at least 2,000 patients and their therapists had appeared on websites on the dark web. Due to bankruptcy, Vastaamo ceased to operate on March 1, 2021. As of February 2022, Finnish policy had recorded around 100 instances of re-victimisation, including fraudulent use of the victims' leaked information.

Taking into account the context in which this hack occurred, the societal impact must be considered as well. Finland is a country that has tried to reduce the stigma around mental health, and encourages its citizens to access help without fear of repercussions. This hack has left people justifiably worried about their own security and privacy, and that of their loved ones. As a result, mental health and victim support charities have been overwhelmed with calls from distressed people who fear that their personal records have been accessed and possibly released to the public.

Attribution

This event has been attributed to the Finnish hacker Julius Kivimäki.

Public attribution was made for this incident. The Finnish criminal investigation has made a legal attribution of the incident to Kivimäki. He has been on remand since February 2023.

A non state actor. Kivimäki is charged with aggravated computer breach, aggravated attempted extortion, aggravated dissemination of information violating personal privacy, extortion, attempted extortion, computer breach, message interception and falsification of evidence.

References

[1] Such as Artificial Intelligence.

[2] WHO definition of health as: "A state of complete physical, mental and social well being and not merely the absence of disease."

[3] The term cyberattack is used more broadly than the meaning of the term under International Humanitarian Law (IHL), which applies only to armed conflicts, and where "attack" has very specific meanings.

[4] Namely cyber incidents below the threshold of armed conflict but beyond regular peacetime relations.

[5] Pursuant to paragraph 3 of General Assembly resolution 73/266.

[6] UN GGE Report, 2021, para 83

[7] UN GGE Report, 2021, para, 85.

[8] https://documents-dds-ny.un.org/doc/UNDOC/GEN/N15/457/57/PDF/N1545757.pdf?OpenElement

[9] The UN norms of responsible state behaviour in cyberspace https://documents.unoda.org/wp-content/uploads/2022/03/The-UN-norms-of-responsible-state-behaviour-in-cyberspace.pdf

[10] Unpacking Cyber Capacity-Building Needs Part II. Introducing a Threat-Based Approach, UNIDIR, Authors Samuele Dominioni and Giacomo Persi Paoli, p.15-16. https://unidir.org/publication/unpacking-cyber-capacity-building-needs-part-ii-introducing-a-threat-based-approach/

[11] Ibid p.37

[12] Ibid , p.40-41

[13] A Taxonomy of Malicious ICT Incidents, United Nations Institute for Disarmament Research (UNIDIR), Dr Samuele Dominioni, and Dr Giacomo Persi Paoli, 2022.

[14] Ibid, p.7

[15] Ioannis Agrafiotis, et al. 2016. "Cyber Harm: Concepts, Taxonomy and measurement" (August 1, 2016). Saïd Business School WP 2016-23.

[16] Further to the Expert Meeting, and building on those discussions, it was considered useful to draw out the potential harmful impact on humanity of information technology / digital systems.

[17] Such as Artificial Intelligence.

[18] https://www.who.int/about/governance/constitution

[19] Council of Europe Cybercrime Convention Committee Mapping Study on Cyber violence, which defined cyber violence as: "The use of computer systems to cause, facilitate, or threaten violence against individuals, that results in (or is likely to result in) physical, sexual, psychological or economic harm or suffering and may include the exploitation of the individual's circumstance, characteristics or vulnerabilities." https://www.coe.int/en/web/cyberviolence/cyberviolence-at-a-glance#:~:text=Cyberviolence%20is%20the%20use%20of,individual's%20circumstances%2C%20characteristics%20or%20vulnerabilities.

[20] Dr Nils Melzer, former United Nations Special Rapporteur on Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment, stated in 2020 that:"[c]ybertechnology can be used to inflict, or contribute to, severe mental suffering while avoiding the conduit of the physical body, most notably through intimidation, harassment, surveillance, public shaming and defamation, as well as appropriation, deletion of manipulation of information."

[21] Defined in the World Health Organization (WHO) Constitution. https://www.who.int/about/accountability/governance/constitution

[22] https://apps.who.int/iris/bitstream/handle/10665/42495/9241545615_eng.pdf. The Violence Prevention Alliance (VPA) addresses the problem of violence as defined in the World report on violence and health (WRVH). https://www.who.int/groups/violence-prevention-alliance/approach

[23] See video: “A Theory of Violence” Robin Coupland, Daniel Dobos and Nathan Taback 2022 https://www.youtube.com/watch?v=SKEXP5Csuyg. R. Coupland, “Armed violence”, Medicine and Global Survival, vol. 7, 2001, pp. 33-37. Taback, Nathan; Coupland, Robin (2005). "Towards Collation and Modelling of the Global Cost of Armed Violence on Civilians". Medicine, Conflict and Survival. 21 (1): 19–27. doi:10.1080/13623690420 00315032. PMID 15690624. S2CID 35779029.

[24] The Theory of Violence assumes that violence can always be expressed in terms of its impact on the victim’s health e.g. lethality, number of people killed, injured, displaced, assaulted, etc. The determinants of the impact of any act of violence in any context are:

- the intent of the perpetrator to cause the impact;
- the physical capacity of the perpetrator for violence (given by the weapons available to cause the impact in question);
- the vulnerability of the victim or victims (given by the potential of the victim to suffer the impact in question.)

In public health terms, the determinants are “risk factors” for the impact in question. If one of the risk factors does not exist, there can be no impact. Any preventive measure is related to one or more of the risk factors. No preventive measure is unrelated to these risk factors. “Harm-oriented” analysis starts with understanding the impact; this permits conclusions to be drawn about the perpetrator’s intent and physical capacity and, most importantly, the victims’ vulnerability.

[25] Translating this Theory of Violence into an analytical tool for cyberattacks requires replacing “physical capacity” with “capacity for cyberattack” that encompasses the capacity of the computer systems available to the perpetrator and the technical expertise to launch an attack.

[26] This is because of the interconnected nature of infrastructure, the inherently dual nature of infrastructure (which is targetable if falling under the definition of a military objective under IHL), and the difficulty to assess the impact and unintended consequences of attacks using cyber tools.

[27] UNIDIR, 2023 “AI and International Security - Understanding the Risks and Paving the Path for ConfidenceBuilding Measures”

[28] Ioannis Agrafiotis, et al. 2016. “Cyber Harm: Concepts, Taxonomy and measurement” (August 1, 2016). Saïd Business School WP 2016-23. University of Oxford. 2016. http://dx.doi.org/10.2139/ssrn.2828646

[29] CyberGreen, “A Cyber Belief Model.” CyberGreen. 2023. https://cybergreen.net/technical-report-23-01/

[30] The Cybergreen belief model emphasizes resilience with regards to impact, as opposed to discussing harm / harm categories specifically.

[31] Powers, Edward W., Fancher, Donald., & Sibler, Justin. “Beneath the surface of a cyber attack”. Deloitte. 2016. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/risk/us-risk-beneath-the-surface-of-a-cyber-attack.pdf

[32] Gisel, Laurent & Olejnik, Lukas. “THE POTENTIAL HUMAN COST OF CYBER OPERATIONS” ICRC Expert Meeting Report. 2018. https://www.icrc.org/en/document/potential-human-cost-cyber-operations

[33] Shandler et al. “Cyberattacks, Psychological Distress, and Military Escalation: An Internal Meta-Analysis”, Journal of Global Security Studies, 8(1), 2023, https://doi.org/10.1093/jogss/ogac042

[34] Dominioni, Samuele & Paoli, Giacomo Persi. “Taxonomy of malicious ICT incidents.” UNIDIR. 2022. https://unidir.org/publication/taxonomy-malicious-ict-incidents

[35] Florian J. Egloff and James Shires, 2023, “The better angels of our digital age? Offensive cyber capabilities and state violence”, European Journal of International Security, first published online in 2021.

[36] Florian J. Egloff and James Shires, 2023, “The better angels of our digital age? Offensive cyber capabilities and state violence”, p. 138-139

[37] 2016, Christine Izuakor , Department of Computer Science, University of Colorado, Understanding the Impact of Cyber Security Risks on Safety

[38] It is difficult to assemble a list of indicators that is exhaustive or sufficient. Even the highly quantitative frameworks, like Deloites' 14 impact factors, at times defer to best guesses or "reliance on assumptions,"# and Agrafiotis et al. note that more research is needed to assess metrics that may be considered as sufficient for given harm types. Agrafiotis et al. 2016. 37

[39] Viasat is an American communications company based in Carlsbad, California, with additional operations across the United States and worldwide. Viasat is a provider of high-speed satellite broadband services and secure networking systems covering military and commercial markets.

[40] After careful consideration, we opted against providing a specific definition for cyber harm. This decision was influenced by the insights shared by experts throughout the workshop, who emphasized that the potential for causing harm, whether through cyber or kinetic methods, can be similar in terms of the types and categories of harm inflicted.

[41] Declaration of Basic Principles of Justice For Victims of Crime and Abuse of Power

[42] Basic Principles and Guidelines on the Right to a Remedy and Reparation for Victims of Gross Violations of International Human Rights Law and Serious Violations of International Humanitarian Law

[43] Attribution is a technical step in international law for attaching a given act or omission to a relevant actor, such as a state, for the purposes of determining who is responsible for a violation of international law and which is the appropriate legal framework establishing the rights and obligations of states affected by an incident and impose consequences, if appropriate. Thus, attribution is critical for accountability under the law, remedy, and redress for victims of cyberattacks. Attribution of an attack under international law triggers State responsibility, and may trigger the application of IHL, and/or activate the right to respond in self defense. An International Armed Conflict (IAC) is triggered "whenever there is a resort to armed force between States ", this definition is "generally considered as the contemporary reference for any interpretation of the notion of armed conflict under humanitarian law". This includes situations in which an (i) act of violence, is attributable to one State, (ii) against the population, armed forces or territory of another State. The existence of an armed conflict must be deduced from the facts. See para. 241 of the 2020 edition of the ICRC Updated Commentary on common Article 2: ICRC Database, Treaties, States Parties and Commentaries, Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977. Article 37 - Prohibition of perfidy, https://ihl-databases.icrc.org/en/ihl-treaties/api-1977/article-37?activeTab=1949GCs-APs-and-commentaries (Last accessed on 22.02.2023).

[44] The challenges and complexity in the attribution of cyberattacks to an actor(s) can be summarized in a three-tiered approach: technical, political, and legal:

Technical attribution - the forensic investigation of a malicious incident to the origins of an attack platform, specific software, hardware, code, or modus operandi.

Political attribution - determining or disclosing by a State who is the party(s) responsible for an attack including a nation state, State-sponsored group, criminal group, collective, etc. based on analysis, assessment and/or judgement.

Legal attribution - determining who is responsible for an attack based on technical means to identify the origin of the attack and legal criteria in order to ascribe legal consequences and/or other sanctions, for example through a court of law. Attribution of a cyberattack under international law may trigger the application of IHL, State responsibility, and/or activate the right to respond in self-defense.

[45] See discussion of the risk of conceptual expansion in Florian J. Egloff and James Shires research article, "The better angels of our digital age? Offensive cyber capabilities and state violence" published in the European Journal of International Security (2023) and first published online in 2021. Published when Egloff was at the Center for Security Studies, ETH Zurich, and Shires was at the Institute for Security Studies and Global Affairs, Leiden University.

[46] Florian J. Egloff and James Shires, "The better angels of our digital age? Offensive cyber capabilities

and state violence" published in the European Journal of International Security (2023) and first published online in 2021, p. 145-6

[47] The paper defines OCCs as "the combination of various elements that jointly enable the adversarial manipulation of digital services or networks. Ibid p. 133.

[48] Jus ad bellum, referring to the conditions under which States may resort to war or to the use of force (UN Charter); International Humanitarian Law (IHL), also known as jus in bello, regulating the conduct of parties engaged in armed conflicts and protecting the victims of such conflicts, including civilians.

[49] Article 35 of Additional Protocol I to the Geneva Conventions prohibits "widespread, long-term and severe damage to the natural environment".

[50] Such as Artificial Intelligence.

[51] WHO definition of health as: "A state of complete physical, mental and social well being and not merely the absence of disease."

[52] For example, emphasis added: "destructive impact on societies as well as individuals" (SG Foreword); "They may also pose direct and indirect harm to individuals." (para. 10); "The COVID-19 pandemic has demonstrated the risks and consequences of malicious ICT activities…." (para. 10); "Norm 13 (a)…… prevent ICT practices that are acknowledged to be harmful or that may pose threats to international peace and security."(para. 19); "Norm 13 (b) In case of ICT incidents, States should consider …. the nature and extent of the consequences". (para. 21); "A State that is victim of a malicious ICT incident should consider…… its scope, scale and impact….., including the incident's bearing on international peace and security…..". (para. 24); "…. State practices such as arbitrary or unlawful mass surveillance may have particularly negative impacts on the exercise and enjoyment of human rights…". (para. 37); "While recognizing the importance of technological innovation….. New and emerging technologies may also have important human rights and ICT security implications…….. guide the development ….in a manner that …..does not negatively impact members of individual communities or groups." (para. 40); "…. possible negative impacts of policies on people…. Norm 13 (f) A State should not conduct or knowingly support ICT activity….. that internationally damages critical infrastructure…" (para 41); "…. CERTs/CSIRTs …. are essential to effectively detecting and mitigating the immediate long term effects of ICT incidents…Harm to emergency response teams can undermine trust and….. And can have wider, often unforeseen consequences…" (para 65); "With regard to this norm, ICT activity that internationally damages critical infrastructure or otherwise impairs the use and operation of critical infrastructure … can have cascading domestic, regional and global effects. It posts an elevated risk of harm to the population…". (para. 42)

[53] "Norm A Consistent with the purposes of the United Nations, including to maintain international peace and security, States should cooperate in developing and applying measures to increase stability and security in the use of ICTs and to prevent ICT practices that are acknowledged to be harmful or that may pose threats to international peace and security." "
Norm B In case of ICT incidents, States should consider all relevant information, including the larger context of the event, the challenges of attribution in the ICT environment, and the nature and extent of the consequences."
Norm F A State should not conduct or knowingly support ICT activity contrary to its obligations under international law that intentionally damages critical infrastructure or otherwise impairs the use and operation of critical infrastructure to provide services to the public.
Norm K. States should not conduct or knowingly support activity to harm the information systems of the authorized emergency response teams (sometimes known as computer emergency response teams or cybersecurity incident response teams) of another State. A State should not use authorized emergency response teams to engage in malicious international activity.

[54] The taxonomy cites Ioannis Agrafiotis et al. 2018. "A Taxonomy of Cyber-Harms: Defining the Impacts of Cyber-Attacks and Understanding How They Propagate". Journal of Cybersecurity, vol. 4, no. 1, 2018.

[55] https://www.icrc.org/en/publication/potential-human-cost-cyber-operations

[56] Ibid, Foreword.

[57] Ibid, Executive Summary.

[58] Ibid, p. 33

[59] https://www.academia.edu/81096454/Cyber_Harm_Concepts_Taxonomy_and_Measurement

[60] 2016, Journal of Clinical Nursing Volume 25, Issue 21-22

[61] https://www.oecd.org/dac/peer-reviews/Development-Results-Note.pdf, p.3

[62] U.S. Code § 1589 - Forced labor

[63] https://www.unodc.org/documents/Cybercrime/AdHocCommittee/Second_session/Documents/Letter_from_AHC_Chair_-_2nd_session_methodology_and_guiding_questions4115.pdf

[64] para. 5: States Party may require the intent to cause harm before criminal liability attaches.

[65] ICRC Human Cost report p. 39 https://www.icrc.org/en/publication/potential-human-cost-cyber-operations

[66] United Nations Environment Programme

[67] https://www.ohchr.org/en/instruments-mechanisms/instruments/declaration-basic-principles-justice-victims-crime-and-abuse

[68] https://www.ohchr.org/en/instruments-mechanisms/instruments/basic-principles-and-guidelines-right-remedy-and-reparation

[69] ICC, Rules of Procedure and Evidence, Rule 85

[70] e.g. Denmark, France, Germany; Kjelgaard, Jeppe Mejer, and Ulf Melgaard. "Denmark's Position Paper on the Application of International Law in Cyberspace: Introduction". Nordic Journal of International Law 92.3 (2023): 446-455. https://doi.org/10.1163/15718107-20230001; OEWG on ICT, International Law Applies to Operations in Cyberspace submitted by France, 1st December 2021, p.3, https://documents.unoda.org/wp-content/uploads/2021/12/French-position-on-international-law-applied-to-cyberspace.pdf
UNGA, Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by States submitted by participating governmental experts in the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security established pursuant to General Assembly resolution, Germany, 13 July 2021, A/76/136, p.35-36, https://front.un-arm.org/wp-content/uploads/2021/08/A-76-136-EN.pdf

[71] As part of the OEWG process to clarify the application of law, Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, A/76/135, 14 July 2021, para 95b.

[72] OEWG on ICT, Position on the application of international law in cyberspace submitted by Poland, 28 July 2023, p.6, https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-_(2021)/The_Republic_of_Poland%E2%80%99s_position.pdf

[73] OEWG on ICT, Position Paper on the Application of International Law in Cyberspace submitted by Costa Rica, 21 July 2023, p.9-10, https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-_(2021)/Costa_Rica_-_Position_Paper_-_International_Law_in_Cyberspace.pdf

[74] OEWG on ICT, International Law Applies to Operations in Cyberspace submitted by France, 1st December 2021, p.4, https://documents.unoda.org/wp-content/uploads/2021/12/French-position-on-international-law-applied-to-cyberspace.pdf

[75] UNGA, Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by States submitted by participating

governmental experts in the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security established pursuant to General Assembly resolution, Australia, 13 July 2021, A/76/136, p.11, https://front.un-arm.org/wp-content/uploads/2021/08/A-76-136-EN.pdf

[76] Ibid, Norway, 13 July 2021, A/76/136, p.70, https://front.un-arm.org/wp-content/uploads/2021/08/A-76-136-EN.pdf

[77] OEWG on ICT, Perspective on Key Legal and Practical Issues Concerning the Application of International Law to Cyber Operations submitted by Israel, 9 December 2020.

[78] For example, in New Zealand's view, States may take into account both the immediate impacts and the intended or reasonably expected consequential impacts. (OEWG on ICT, The Application of International Law to State Activity in Cyberspace submitted by New Zealand, 24 July 2023, para 7

[79] OEWG on ICT, National position on international law and cyberspace submitted by Finland, 24 July 2023, p.6, https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-_(2021)/Finland_position_IL_cyberspace.pdf;
In Denmark's view, economic or political coercion can qualify as a use of force (Kjelgaard, Jeppe Mejer, and Ulf Melgaard. "Denmark's Position Paper on the Application of International Law in Cyberspace: Introduction". Nordic Journal of International Law 92.3 (2023): 446-455. https://doi.org/10.1163/15718107-20230001)

[80] France considers that it may amount to a prohibited use of force (OEWG on ICT, International Law Applies to Operations in Cyberspace submitted by France, 1st December 2021, p.3, https://documents.unoda.org/wp-content/uploads/2021/12/French-position-on-international-law-applied-to-cyberspace.pdf)
Ireland considers that the illegal use of force may include instances where a cyber operation does not cause physical damage, but significant impairment of functionality of critical infrastructure (OEWG on ICT, Position Paper on the Application of International Law in Cyberspace by Ireland, 7 July 2023, para 18 https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-_(2021)/Ireland_-_National_Position_Paper.pdf)
Italy considers it controversial, unless there is interruption of essential services, without physical damage (OEWG on ICT, Italian Position Paper on International Law and Cyberspace, 1 December 2021, p.8, https://documents.unoda.org/wp-content/uploads/2021/10/italian-position-paper-international-law-and-cyberspace.pdf)

[81] Military and Paramilitary Activities in and against Nicaragua (Nicaragua v United States of America), Judgment, 1986 ICJ Reports 14 ('Nicaragua case'), para 191.

[82] Military and Paramilitary Activities in and against Nicaragua (Nicaragua v United States of America), Judgment, ICJ Report 1986, para 195.

[83] Advisory Opinion of the International Court of Justice on the Legality of the Threat or Use of Nuclear Weapons, ICJ Rep. 1996, para 41.

[84] For example, Ireland holds the view that loss or impairment of functionality to ICT infrastructure without physical damage may amount to an armed attack if it is inflicted on such a scale and with such effects that it is comparable to a conventional armed attack. (OEWG on ICT, Position Paper on the Application of International Law in Cyberspace by Ireland, 7 July 2023, para 3, https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-_(2021)/Ireland_-_National_Position_Paper.pdf)
Singapore holds the same view, e.g. targeted cyber operation causing sustained and long-term outage of Singapore's critical infrastructure (UNGA, Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by States submitted by participating governmental experts in the Group of Governmental Experts on

Advancing Responsible State Behaviour in Cyberspace in the Context of International Security established pursuant to General Assembly resolution, Singapore, p. 84, para 8, https://front.un-arm.org/wp-content/uploads/2021/08/A-76-136-EN.pdf)

[85] For example, France shares this view (OEWG on ICT, International Law Applies to Operations in Cyberspace submitted by France, 1st December 2021, p.7, https://documents.unoda.org/wp-content/uploads/2021/12/French-position-on-international-law-applied-to-cyberspace.pdf)
Singapore also shares this view (UNGA, Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by States submitted by participating governmental experts in the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security established pursuant to General Assembly resolution, Singapore, 13 July 2021, A/76/136, p. 84, para 9, https://front.un-arm.org/wp-content/uploads/2021/08/A-76-136-EN.pdf)

[86] For example, Poland, Denmark, Germany, Netherlands and the USA share the view that the right to self-defence applies in this case: OEWG on ICT, Position on the application of international law in cyberspace submitted by Poland, 28 July 2023, p.6; Kjelgaard, Jeppe Mejer, and Ulf Melgaard. "Denmark's Position Paper on the Application of International Law in Cyberspace: Introduction". Nordic Journal of International Law 92.3 (2023): 446-455. https://doi.org/10.1163/15718107-20230001; UNGA, Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by States submitted by participating governmental experts in the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security established pursuant to General Assembly resolution, 13 July 2021, A/76/136, p.43, p.65, p.137 https://front.un-arm.org/wp-content/uploads/2021/08/A-76-136-EN.pdf)
France and Brazil do not support this view: OEWG on ICT, International Law Applies to Operations in Cyberspace submitted by France, 1st December 2021, p.7, https://documents.unoda.org/wp-content/uploads/2021/12/French-position-on-international-law-applied-to-cyberspace.pdf; UNGA, Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by States submitted by participating governmental experts in the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security established pursuant to General Assembly resolution, Brazil, 13 July 2021, A/76/136, p.20 https://front.un-arm.org/wp-content/uploads/2021/08/A-76-136-EN.pdf)

[87] Australia allows itself 'anticipatory self-defence'. (UNGA, Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by States submitted by participating governmental experts in the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security established pursuant to General Assembly resolution, Australia, 13 July 2021, A/76/136, p.6, https://front.un-arm.org/wp-content/uploads/2021/08/A-76-136-EN.pdf)
France allows itself to use pre-emptive self-defence in response to a cyberattack that "has not yet been triggered but is about to be, in an imminent and certain manner, provided that the potential impact of such an attack is sufficiently serious". (OEWG on ICT, International Law Applies to Operations in Cyberspace submitted by France, 1st December 2021, p.7, https://documents.unoda.org/wp-content/uploads/2021/12/French-position-on-international-law-applied-to-cyberspace.pdf)
Germany and Brazil are opposed to the use of preventive self-defence (UNGA, Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by States submitted by participating governmental experts in the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context

of International Security established pursuant to General Assembly resolution, 13 July 2021, A/76/136, p.20, p.43, https://front.un-arm.org/wp-content/uploads/2021/08/A-76-136-EN.pdf)

[88] Kjelgaard, Jeppe Mejer, and Ulf Melgaard. "Denmark's Position Paper on the Application of International Law in Cyberspace: Introduction". Nordic Journal of International Law 92.3 (2023): 446-455. https://doi.org/10.1163/15718107-20230001.

[89] OEWG on ICT, International Law Applies to Operations in Cyberspace submitted by France, 1st December 2021, pp 5-6, https://documents.unoda.org/wp-content/uploads/2021/12/French-position-on-international-law-applied-to-cyberspace.pdf

[90] OEWG on ICT, The Application of International Law to State Activity in Cyberspace submitted by New Zealand, 24 July 2023, para 8, https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-_(2021)/The_Application_of_International_Law_to_State_Activity_in_Cyberspace.pdf

[91] "Sovereignty in the relations between States signifies independence. Independence in regard to a portion of the globe is the right to exercise therein, to the exclusion of any other State, the functions of a State." Island of Palmas (Netherlands v. the US), PCA 1928, 2 UNRIAA 829, 838.

[92] UN Charter, Article 2(1).

[93] In their submissions to the OEWG, as part of the process to clarify the application of law: Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, A/76/135, 14 July 2021, para 95b

[94] OEWG on ICT, Position on the application of international law in cyberspace submitted by Poland, 28 July 2023, p.3, https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-_(2021)/The_Republic_of_Poland%E2%80%99s_position.pdf

[95] OEWG on ICT, National position on international law and cyberspace submitted by Finland, 24 July 2023, p.2, https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-_(2021)/Finland_position_IL_cyberspace.pdf

[96] OEWG on ICT, Position Paper on the Application of International Law in Cyberspace submitted by Costa Rica, 21 July 2023, p.6, para 21, https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-_(2021)/Costa_Rica_-_Position_Paper_-_International_Law_in_Cyberspace.pdf

[97] OEWG on ICT, International Law applicable in cyberspace submitted by Canada, 1 March 2022, https://www.international.gc.ca/world-monde/issues_development-enjeux_developpement/peace_security-paix_securite/cyberspace_law-cyberespace_droit.aspx?lang=eng

[98] UNGA, Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by States submitted by participating governmental experts in the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security established pursuant to General Assembly resolution, Norway, 13 July 2021, A/76/136, p.67, https://front.un-arm.org/wp-content/uploads/2021/08/A-76-136-EN.pdf

[99] See Tallinn Manual 2.0, paras 10-12 on Rule 92, pp. 417-418 referenced in ICRC Expert Meeting Report, 2019, p. 73

[100] The principle of distinction between military objectives and civilians and civilian objects is enshrined in Articles 48, 51(2), 52(2) of API and the rules 1 and 7 of CIHL. It prohibits any indiscriminate attack targeting anything but military objectives as defined under Art. 52(2). Those provisions protect the civilian population or individuals who shall not be the object of an attack "unless and for such time" as they are directly participating in the hostilities, as well as civilian infrastructures unless they turn into a military objective. In addition, dual-use objects used for both for military and civilian purposes are to be

considered as military objectives, even if the military use is secondary. In case of doubt as to whether an object that is normally dedicated to civilian purposes is being used to make an effective contribution to military action, it must be presumed to remain protected as a civilian object.

[101] p.10

[102] Kayishema, Case No. ICTR-95-1-T, para. 10

[103] Akayesu, Case No. ICTR-96-4-T, para. 104

[104] Prosecutor v. Muhimana, Case No. ICTR-95-1B-T, Judgment, para. 502 (Apr. 28, 2005)

[105] Nema Milaninia, Understanding Serious Bodily or Mental Harm as an Act of Genocide, Vanderbilt Journal of Transnationl Law, 2018, p.1395

[106] Digital Front Lines, Technology will not exceed our Humanity, Karim A.A. Khan KC

[107] https://techcrunch.com/2022/03/31/viasat-cyberattack-russian-wiper/

[106] https://www.sentinelone.com/labs/acidrain-a-modem-wiper-rains-down-on-europe/

[109] https://therecord.media/viasat-hack-was-two-incidents-and-resulted-in-sanctions

[110] https://www.state.gov/attribution-of-russias-malicious-cyber-activity-against-ukraine/

[111] https://www.sentinelone.com/labs/acidrain-a-modem-wiper-rains-down-on-europe/

[112] https://www.consilium.europa.eu/en/press/press-releases/2022/05/10/russian-cyber-operations-against-ukraine-declaration-by-the-high-representative-on-behalf-of-the-european-union/

[113]https://www.gov.uk/government/news/russia-behind-cyber-attack-with-europe-wide-impact-an-hour-before-ukraine-invasion

[114] https://www.state.gov/attribution-of-russias-malicious-cyber-activity-against-ukraine/

CyberPeace Institute

The CyberPeace Institute is an independent and neutral nongovernmental organization whose mission is to ensure the rights of people to security, dignity and equity in cyberspace. The Institute works in close collaboration with relevant partners to reduce the harms from cyberattacks on people's lives worldwide. By analyzing cyberattacks, the Institute exposes their human impact, how international laws and norms are being violated, and advances responsible behavior to enforce cyberpeace.

CyberPeace Institute
Campus Biotech Innovation Park
Avenue de Sécheron 15
1202 Geneva, Switzerland

cyberpeaceinstitute.org
@CyberPeace Institute
@CyberpeaceInst