LUISS

Corso di laurea in Data Science and Management

Cattedra Data Privacy and Security

A threat intelligence grading system for the financial sector

Prof. Paolo Spagnoletti

RELATORE

Prof. Irene Finocchi

CORRELATORE

Matr. 745941 - Andrea Buscemi

CANDIDATO

Anno Accademico 2022/2023

Sommario

Preface

On the 24th of February, 2022, the Russian Federation launched its special military operation invading Ukraine, waking all of Europe from its quasi-dormient state it found itself since the end of WW2. The population of Europe (especially in the Western part, such as us here in Italy), was reminded once more that conflicts still arise and, for this reason, you must prepare.

The war in Ukraine has decisively shown that one very important part of modern conflicts is the cyberspace, if were there to be still some skeptics.

The belligerents (and the co-belligerents) are in fact making extensive use of the cyberspace to achieve strategic objectives (undermining population support to the war effort or foreign countries support), operational objectives (sabotage of critical infrastructures) and even tactical objectives (geo-localizing enemy units or targets). Thus, showing the relevance of cyber operations throughout the full conflict spectrum.

The Russian Federation in particular (through third party actors but still attributable to Russian state entities) has used the cyberspace to undermine popular support in foreign countries, including ours, via cyber-attacks on civilian services, such as the attack to banks and companies in February 2023 or the more recent attacks on government websites in August 2023.

The environment we are experiencing thus calls for immediate action and planning towards cyber threats and consequent defense tactics.

I have exchanged views on this subject with several professionals in the field I have encountered in the last year, from different backgrounds and stances (furtherly explained later), and all these interactions have ignited a spark in a specific direction for a deeper study and fresh ideas that might help in our journey, national and intra-allied, to cyber resilience and deterrence.

The interest arose in the field of evaluating intelligence, cyber threats, in particular, focusing on how we could more efficiently evaluate in advance a cyber threat coming from an intelligence source (especially from an open source). A field called “Cyber Threat Intelligence” (CTI).

I will list here the major encounters that spurred the research you are about to read, though many more existed, and all contributed in some way.

The first of these encounters was with a former Italian Army intelligence officer, now an intelligence and security consultant, who focuses on Open-Source Intelligence (OSINT from now on), who

introduced me to the relevance, especially in the Ukraine war case, of the cyber domain and especially of OSINT in terms of counteracting strategic campaigns of misinformation.

The second encounter was with the members of CERTFin, an entity formed by the Italian Central Bank and major private banks, which focuses on the cyber menaces threatening the Italian financial sector. They introduced me to their approach, tools and rationale, and it was with them first that the CTI idea I will expose later came about.

The third and final encounter was during my time participating in the Italian Navy exercise Mare Aperto 23-1, where I had the fortune and honor to meet the many men and women serving our country on the waves. I have been able to exchange the initial ideas I had with several intelligence and counter-intelligence officers, who all gave me their precious perspective and helped me by showing me first-hand how the intelligence evaluation process works, and how could it be expanded.

The idea

The CTI world I came to know, deriving from a military background, and operating in the NATO framework (in 2016 NATO declared the cyberspace an official domain of warfare), derives its processes and practices from military intelligence.

In particular, and limiting ourselves to the focus of this research, to evaluate intel on possible cyber threats we use the same system adopted by NATO intelligence, that is, the NATO Admiralty System.

The system (NATS for short) evaluates two dimensions: credibility of the source (A to F), and credibility of the information (1 to 6), thus creating a matrix useful to decide whether the intel is to be considered reliable, uncertain or unreliable.

If you are evaluating all-encompassing forms of intelligence, it works just fine. But, if you are just evaluating threats, that is, you already know the info is about a vulnerability in a system, you might want to expand the system and add an evaluation of the threat itself.

That's what this research will be about, to expand on the NATO Admiralty System specifically for Cyber Threat Intelligence, so to evaluate not only the reliability of source and information, but also the magnitude of the possible threat on the victim ("impact") and the grade to which the vulnerability is spread across the whole system/industry you are considering ("systemic diffusion").

1 – Introduction

The financial sector is a vital part of the global economy, and in recent decades has become more digitalized and interconnected than ever. The financial sector is needed for the functioning of the real economy, as it must perform a variety of key functions reliably and efficiently. Payment services, securities trading, settlement services, deposits, lending and many others.

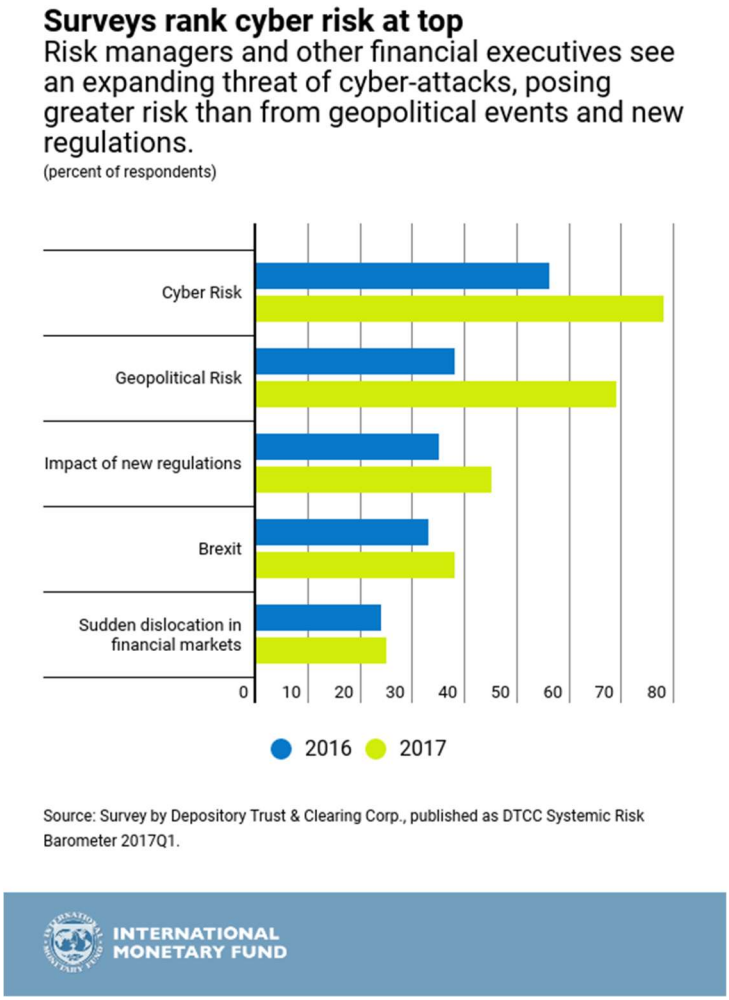
All these processes have become increasingly digitalized, creating new and important interdependencies, that's why the financial sector has come to rely on a robust information and communication technology infrastructure, to ensure confidentiality, integrity and availability of data and systems. It is therefore self-evident that cyber threats could disrupt, compromise and critically impale the information systems and data of financial institutions and actors worldwide.

The interconnectedness of information systems allows for cyber threats to spread immediately and widely as never seen before with other kinds of threat, aided by the speed at which a cyber incident occurs. Malicious threats are also becoming more persistent and prevalent, manifesting the high level of coordination and sophistication achieved by cyber threat actors.

According to the Cybersecurity Ventures 2022 Official Cybercrime Report, the global annual cost of cybercrime was projected to reach $8 trillion annually. If cybercrime were to be a country, it would be the third economy in the world by GDP, behind the US and China, in front of Japan and Germany.

Not surprisingly, surveys consistently show that risk managers and other executives at financial institutions worry most about cyber-attacks, as in the graph below elaborated by the International Monetary Fund shown in a 2018 article, written by the now President of the European Central Bank, Christine Lagarde.

Figure 1 - Surveys rank cyber risk at top - IMF, 2017



Cyber risk can be defined as "operational risks to information and technology assets that have consequences affecting the confidentiality, availability, or integrity of information or information systems" (Cebula and Young, 2010).

Compared to risk categories covered by insurance, cyber risk shares characteristics with both property and liability risk, as well as catastrophic and operational risk (Eling and Wirfs, 2016). On the one hand, cyber risk can impact first (the target) and third parties (a counterpart to the target). On the other hand, losses due to cyber risk are frequently small and independent but they could also have a low frequency and a high impact ('blackout scenario'). Cyber risk can be unrelated to cyberattacks for example, software updates or natural disasters can lead to the crystallization of cyber risk through business disruptions without any nefarious intent, as outlined in the definition of cyber incidents.

Cyber-attacks can impact firms through the three main aspects of information security: confidentiality, integrity and availability. Confidentiality issues arise when private information within a firm is disclosed to third parties as in the case of data breaches. Integrity issues relate to misuse of the systems, as is the case for fraud. 3 Finally, availability issues are linked to business disruptions. The three types of cyber-attacks have different direct impacts on the targets: Business disruptions prevent firms from operating, resulting in loss revenue; fraud leads to direct financial losses; while the effects of data breaches take more time to materialize, through reputational effects as well as litigation costs.

As the US Federal Reserve Board of Governors put it in their July 2022 Report to Congress:

"The rising number of advanced persistent threats increases the potential for malicious cyber activity within the financial sector. These threats may result in incidents that affect one or more participants in the financial services sector simultaneously and have potentially systemic consequences. Such incidents could affect the ability of targeted firms to provide services and conduct business as usual, presenting a unique challenge to operational resilience. These incidents can also threaten the confidentiality, integrity, and availability of the targeted firm's data."

It is therefore perfectly clear that the cyber threat to the financial sector is a matter of the utmost priority and importance.

A fundamental part of counteracting a threat, of whatever type it might be, physical or digital, is **knowing** the threat itself. That's why intelligence services exist, the more you know about what you must go up against, the better you can prepare and act when the time comes.

In the field of cyber threat intelligence, much work is still to be done, and this work aims at providing an evaluative system, capable of overcoming the limitations presented by what is currently being used in the field, an intelligence evaluation system taken from military intelligence, and expanding on it. In the hope that it will help cybersecurity decision makers prepare better for what awaits them.

We will now go over how intelligence evaluation works and its role in evaluating cyber threats in the financial sector.

1.1 - The Intelligence Cycle

The Intelligence cycle, as defined by the NATO Document AJP-2 (Allied Joint Doctrine for Intelligence, Counterintelligence and Security) (NATO, 2016) *is the sequence of activities whereby information is obtained, assembled, converted into intelligence and made available for users*.

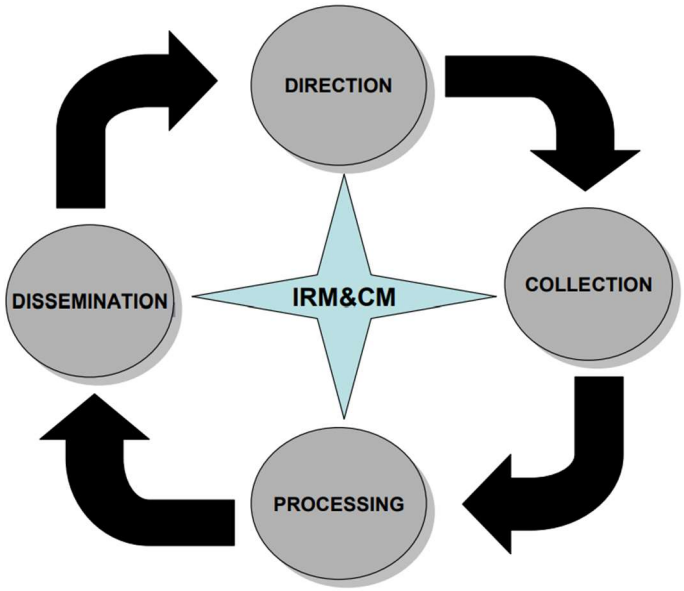
These activities are carried out in the 4 main stages of the intelligence cycle, many different approaches exist in the literature, dividing the cycle in more or fewer steps, in this analysis we'll adopt the NATO standard.

The process is a cycle because it identifies intelligence gaps, unanswered questions, which prompt new collection requirements, thus restarting the intelligence cycle. Intelligence analysts identify intelligence gaps during the analysis phase. Intelligence analysts and consumers determine intelligence gaps during the dissemination and re-evaluation phase.

The 4 stages are identified as:

1. Direction
    1. The *determination of collection requirements, planning the collection efforts, issuing of orders and requests to collection agencies, and maintenance of a continuous check on the productivity of such agencies.*
2. Collection
    1. *The exploitation of sources by collection agencies and the delivery of the information obtained to the appropriate processing unit for use in the production of intelligence.*
3. Processing
    1. *The conversion of information into intelligence through collation, evaluation, analysis, integration and interpretation. 57 Processing is iterative and may generate further requirements for collection before dissemination of the intelligence.*
4. Dissemination
    1. *The timely conveyance of intelligence, in an appropriate form and by any suitable means, to those who need it. It also requires security, conformity to the customer's requirement and a mechanism for feedback.*

Figure 2 - Graphical representation of the intelligence cycle - AJP2



1.2 - Focus: Processing phase

The processing phase entails a structured series of activities which, although set out sequentially, may also occur concurrently. Processing is conducted at several points within the intelligence function. It is a multi-faceted phase of the Intelligence cycle consisting of:

1. Collation: grouping together related items of information or intelligence and provides a record of events, which facilitates further processing.
2. **Evaluation:** the appraisal of an item of information in respect to the reliability of the source and the credibility of the information (this is what we'll try to expand and hopefully improve).
3. Analysis: the information is subjected to review in order to identify significant facts for subsequent interpretation.
4. Integration: analyzed information and/or intelligence is selected and combined into a pattern during the production of further intelligence
5. Interpretation: the significance of information or intelligence is judged in relation to current knowledge.

1.3 - Deeper Focus: Evaluation (NATO Admiralty System)

For several reasons, either because of deception, subjectivity or others, an information might be unreliable, that's why there is an evaluation phase in the processing stage.

Evaluation is the second step in the processing phase and consists of the appraisal of an item of information in respect to the reliability of the source and the credibility of the information.

The evaluation system is called NATO Admiralty System (or code).

An alphanumeric grade is assigned to every piece of information or intelligence, and it tells the degree of assurance on the information or intel itself.

Two dimensions are graded:

- Reliability (of the source): graded from A to F.
- Credibility (of the information): graded from 1 to 6.

Figure 3 - NATO Admiralty System - AJP2

	Reliability of the source		Credibility of the information
A	Completely reliable	1	Confirmed by other sources
B	Usually reliable	2	Probably true
C	Fairly reliable	3	Possibly true
D	Not usually reliable	4	Doubtful
E	Unreliable	5	Improbable
F	Reliability cannot be judged	6	Truth cannot be judged

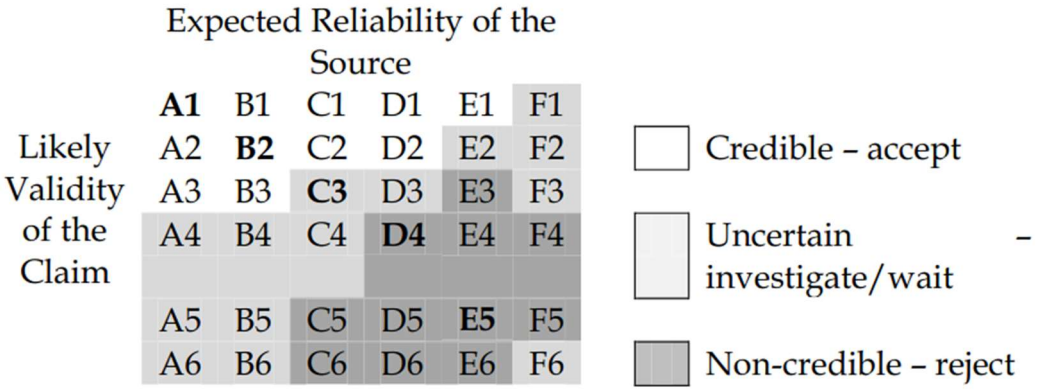
The grade is determined partly by the experience of other information derived from the same source (in case of a sensor, by the known accuracy), partly by the subjective judgement of the evaluator.

The two dimensions are independent of each other, to not have the reliability of the source negatively (or positively) influence the credibility of the information and vice versa.

Not every information produced by a reliable source is accurate, and not every correct information demonstrates the reliability of a source.

The evaluation results then in a matrix capable of providing valuable insight on a piece of intel.

Figure 4 - NATO Admiralty System evaluation matrix - AJP2



1.4 - Open-Source Intelligence (OSINT)

We'll now go through a quick overview of OSINT, a not so new branch of intelligence, but that has gained an undeniable popularity in recent times.

OSINT is a branch of intelligence, tasked with research, collection and analysis of data and news of public interest, deriving from open sources.

The origin of OSINT can be traced back to the exploitation of available information, spoken or written, to plan investments, either of civilian or military scope. A classic example is that of the Lloyds in London, Edward Lloyd's "Coffee House" was in fact a place thoroughly visited by sailors, merchants and shipowners, who would meet there and exchange information on their business, and from this the world's most famous insurance company had its birth.

The need for information, prior to committing to an investment, allowed for the rise of the a "talker" figure, basically an agent that would collect information and would then offer an intelligence service to those who subscribed to its services, the first intelligence/news agency.

Famous Italian examples are Giovanni Poli (Rome), Giovanni Sabadino degli Arienti (Bologna) and Benedetto Fei (Florence), who, in the XV and XVI centuries, collected information and then sent dispatches through all of Europe.

In a stricter military intelligence sense, the birth of OSINT can be traced back to the start of World War Two, when the British Government, through the BBC, set up the BBC Monitoring Service, still active today.

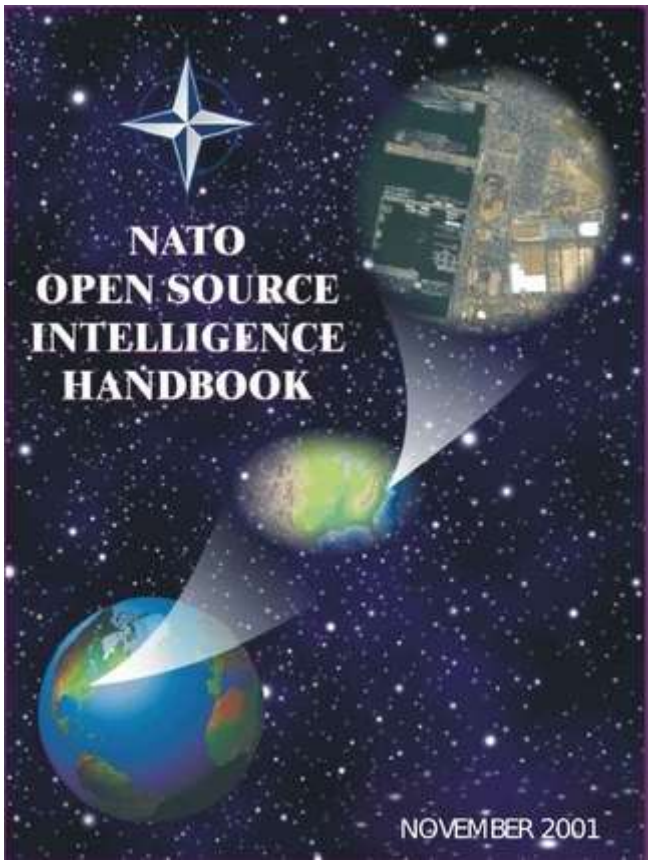
The BBC Monitoring Service was a listening room, where hundreds of “monitors” were employed, mostly refugees, listening to radio and media production worldwide 24 hours a day, in every European language. Winston Churchill used to call the offices in the middle of the night asking (about Hitler): “What's that fellow been saying?”.

During WW2 OSINT was extensively used, and of course later during the Cold War, leading to today’s NATO doctrine and organization about it.

As per the NATO OSINT Handbook (NATO, 2001), Open-Source Intelligence, or OSINT, *is unclassified information that has been deliberately discovered, discriminated, distilled and disseminated to a select audience in order to address a specific question. It provides a very robust foundation for other intelligence disciplines. When applied in a systematic fashion, OSINT products can reduce the demands on classified intelligence collection resources by limiting requests for information only to those questions that cannot be answered by open sources.*

Open information sources are not the exclusive domain of intelligence staffs. Intelligence should never seek to limit access to open sources. Rather, intelligence should facilitate the use of open sources by all staff elements that require access to relevant, reliable information. Intelligence staffs should concentrate on the application of proven intelligence processes to the exploitation of open sources to improve its all-source intelligence products. Familiarity with available open sources will place intelligence staffs in the position of guiding and advising other staff elements in their own exploitation of open sources.

Figure 5 - NATO OSINT Handbook Cover - NATO



OSINT has been on the rise in the recent years, the latest studies estimate that OSINT now makes up between 70 and 90 percent of all intelligence material (Unver, 2018).

Moreover, the Ukraine conflict has proven how valuable OSINT can be even to troops on the field, allowing for geo-targeting and geo-localization of even the smallest units (one of the most brilliant examples being the live map updated by the Italian historian Mirko Campochiari[1].).

During the conflict, OSINT has been also used to debunk and counteract misinformation efforts, one of the most notable examples being the Bellingcat investigation debunking Russian claims about the staging of the Bucha massacre.

1 https://geo.parabellumthinktank.com/index.php/view/map?repository=contemporary&project=real_time_map_russian_ukrainian_conflict_updates; Last accessed 27/09/2023.

1.5 - Introduction to Cyber Threat Intelligence

Cyber Threat Intelligence is what information about cyber threats becomes once it is collected, evaluated and analyzed. Like all intelligence, its usefulness lies in providing a value-add to information on cyber threat, reducing uncertainty for the customer, aiding him in identifying threats and opportunities. To produce accurate, timely and relevant intelligence, the analyst must identify similarities and differences in vast quantities of data and information and detect deceptions.

Just as regular intelligence, as we outlined earlier, it follows a cycle, the intelligence cycle: requirements are defined, data collection is planned, then implemented and evaluated, results are analyzed to produce intelligence, and the resulting intelligence is disseminated and re-evaluated with new information and consumer feedback.

Figure 6 - The Cyber Threat Intelligence cycle

In cyber threat intelligence, analysis often hinges on the triad of actors, intent, and capability, with consideration given to their tactics, techniques, and procedures (TTPs), motivations, and access to the intended targets. By studying this triad, it is often possible to make informed, forward-leaning strategic, operational, and tactical assessments.

Strategic intelligence assesses disparate bits of information to form integrated views. It informs decision and policy makers on broad or long-term issues and/or provides a timely warning of threats. Strategic cyber threat intelligence forms an overall picture of the intent and capabilities of malicious cyber threats, including the actors, tools, and TTPs, through the identification of trends, patterns, and emerging threats and risks, in order to inform decision and policy makers or to provide timely warnings.

Operational intelligence assesses specific, potential incidents related to events, investigations, and/or activities, and provides insights that can guide and support response operations. Operational or technical cyber threat intelligence provides highly specialized, technically focused, intelligence to guide and support the response to specific incidents; such intelligence is often related to campaigns, malware, and/or tools, and may come in the form of forensic reports.

Tactical intelligence assesses real-time events, investigations, and/or activities, and provides day-to-day operational support. Tactical cyber threat intelligence provides support for day-to-day operations and events, such as the development of signatures and indicators of compromise (IOC). It often involves limited application of traditional intelligence analysis techniques.

Figure 7 - Strategic, Operational and Tactical Level in Cyber Threat Intelligence

Properly applied cyber threat intelligence can provide greater insight into cyber threats, allowing for a faster, more targeted response as well as resource development and allocation. For instance,

it can assist decision makers in determining acceptable business risks, developing controls and budgets, in making equipment and staffing decisions (strategic intelligence), provide insights that guide and support incident response and post-incident activities (operational/technical intelligence), and advance the use of indicators by validating, prioritizing, specifying the length of time an indicator is valid (tactical intelligence).

1.6 - Cyber Threat Intelligence in the financial sector

The financial sector is one of the most vulnerable and attacked in terms of cyber events and cyber threats, globally, cyber-attacks directed towards financial institutions are growing exponentially, either be it in terms of sheer number or in terms of dangerousness (more recent trend).

Looking at the numbers, the Clusit (Italian Association for ICT security) 2022 report on Cybersecurity (Clusit, 2022) paints an unconvertible picture, globally, +8.4% attacks in the first semester of 2022 compared to the same period in 2021. An astonishing average of 190 cyberattacks per month, compared to 171 of the previous year.

The relative weight of cybercrime as purpose of the attacks stands at 78% (86% in 2021), giving way to more Espionage and Sabotage (13%, +2% on 2021) and Information Warfare (5%, +3% on 2021).

In the financial world, cyber-threats, as in other sectors, are continuously evolving.

For example, in the most recent years, in Italy, we are witnessing a trend of more dangerous attacks in terms of quality, albeit being less attacks in terms of quantity.

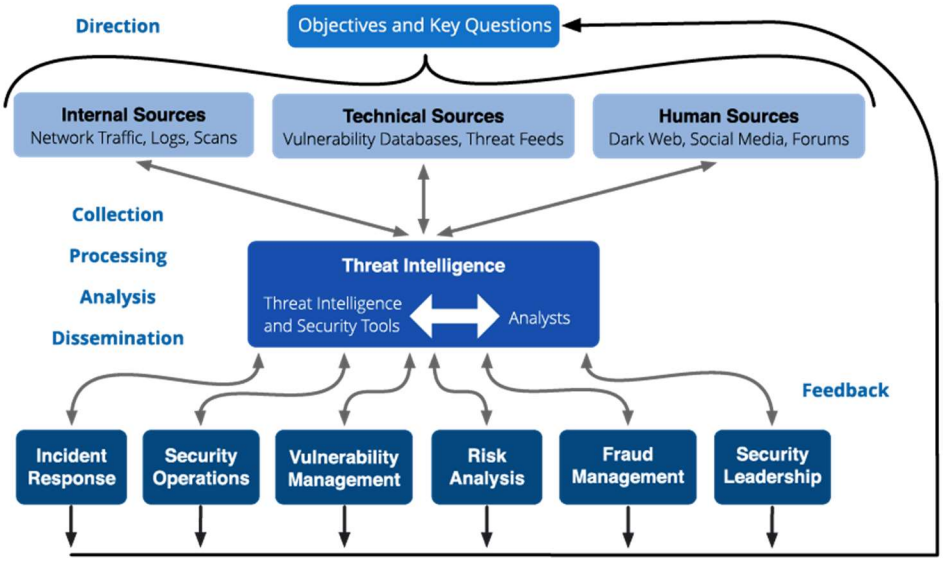
Financial institutions are of course trying to counterattack, introducing new tactics and operational methods capable of aiding in counteracting the ever-evolving cyber-threats.

They are steadily understanding the impelling necessity to move from a classical risk-management approach to a preventive system of dealing with cyber-threats.

This is exactly where Cyber Threat Intelligence comes in, the capability to collect and analyze data, from a variety of different information sources (internal, technical, human or open source), in order to put in place preventive systems of counteraction to possible cyber-threats.

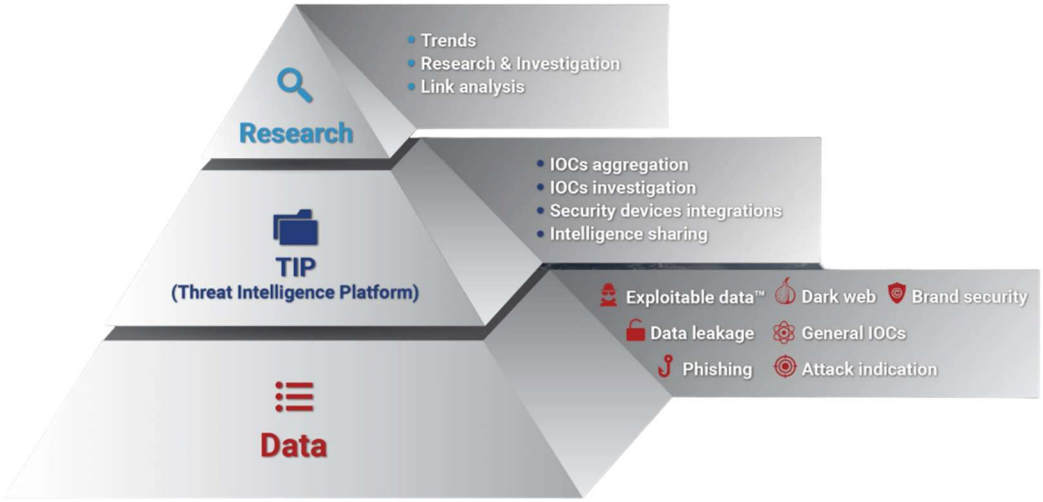
Cyber Threat Intelligence activities are usually carried out by CERTs (Computer Emergency Response Teams), in large businesses, or by CSIRTs (Computer Security Incident Response Teams), usually referring to a government or sectorial agency.

Figure 8 - Assets and Processes of Cyber Threat Intelligence



Their mission is sharing evidence resulting from their analysis with their respective constituencies in order to support CISOs (Chief Information Security Officers) and their operational teams, as well as SOC Providers (Security Operation Center) in some internal activities, such as vulnerability management, risks and threats analysis, fraud and security prevention, crisis management and information sharing.

Figure 9 – The Role of Treat Intelligence Platforms



The Cyber Threat Intelligence cycle of course leverages the latest and most advanced technological solutions in order to better carry out its mission.

One of the most recent and useful tools in Cyber Threat Intelligence today are Threat Intelligence Platforms (TIPs). TIPs are aggregators, used to collate and examine data on cyber threats from multiple and various information sources, and capable of automating some of the steps in the Cyber Threat Intelligence Cycle.

Through automation, integration, standardization, correlation and collaboration capabilities TIPs can exponentially increase the efficiency of Cyber Threat Intelligence activities by improving the relevance, recentness and pregnancy of the generated intel.

2 – Literature Review

We will now cover what the most prominent research has put forward in terms of critiques of the current NATO Admiralty Code, and thus, how to advance it.

Later, we will explore some evaluation systems that could be helpful in terms of creating an expanded grading system, for what concerns the financial sector, mainly moving from the EU Directive PSD2.

2.1 – NATO Comprehensive Review

In June 2020, the NATO Science and Technology Organization published an STO Technical Report, resulting of the research conducted by Research Task Group SAS-114, titled "Assessment and Communication of Uncertainty in Intelligence to Support Decision-Making". Part II, Chapter 7 of the report is a critical examination of the current standards for evaluating source reliability and information credibility (the NATO Admiralty Code), while also highlighting avenues for future research. Research Task Group SAS-114 identified three main sets of issues in the current standards: semantic issues, source reliability determinants and information credibility determinants.

2.1.1 – Semantic Issues

A demonstrably intuitive progression of the qualitative ratings assigned to reliability and credibility is evident in the NATO Admiralty Code (Samet, 1975). Nonetheless, subjective interpretations of the boundaries between these ratings are likely to vary among different analysts, as are the interpretations of the determinant rating criteria (Capet and Revault D'Allones, 2014). In many versions of the NATO System, a reliable "A" source is one possessing a "history of complete reliability," while a usually reliable "B" source has a "history of valid information most of the time" (US Dept. of the Army, 2006; US Dept. of the Army, 2010; US Dept. of the Army, 2011 and US Dept. of the Army, 2012; Canada Dept. of National Defence, 2011).

None of the standards examined associate these qualitative descriptions with a quantitative estimate, a numeric value, (i.e., 'batting averages'), potentially leading to miscommunication.

One analyst may assign usually reliable to sources that provide valid information more than 70% of the time. Another analyst receiving this rating may interpret it to mean that the source provides

reliable information more than 90% of the time and place more confidence in the source than is warranted by the first analyst. On the contrary, an analyst may deem that usually reliable reflects valid information only more than 50% of the time, and prematurely discount the source.

Asked to assign absolute probability values to reliability and credibility ratings, US intelligence officers demonstrated considerable variation in their interpretations (Samet, 1975). For example, probabilistic interpretations of usually reliable and probably true ranged from .55 to .90 and .53 to .90, respectively, while interpretations of fairly reliable and possibly true both ranged from .40 to .80 (Samet, 1975).

Among the military standards considered, reliable or completely reliable indicates the maximum score for source reliability, while confirmed, confirmed by other sources, or completely credible marks the highest degree of information credibility. In spite of these discrepancies, most scales faithfully reproduce the Admiralty Code's A – F (reliability) / 1 – 6 (credibility) scoring system, and ratings are often communicated using only the appropriate alphanumeric code (e.g., A1).

These terminological variations may therefore contribute to miscommunication between users familiar with different standards. For instance, under most US standards examined (US Dept. of the Army, 2006; US Dept. of the Army, 2010a; US Dept. of the Army, 2010b; US Dept. of the Army, 2012), "A" is defined as "reliable", while UK Joint Doctrine 2-00 (UK Ministry of Defence, 2011) defines "A" as "completely reliable" (conforming to NATO doctrine).

A US analyst considering "A" as reliable might transmit that rating to a UK counterpart, who interprets it as completely reliable. This translation is potentially problematic, given that an analyst or consumer may place more weight on a source labelled completely reliable than one labelled reliable. Alternatively, the translation from completely reliable to reliable could lead a recipient to undervalue a source.

Inter-standard miscommunication and misinterpretation could also arise where evaluation scales adopt "accuracy" as a synonym for information credibility (US Dept. of the Army, 2010a and US Dept. of the Army, 2010b). While credibility often includes considerations of accuracy, it is likely a more complex construct, with more than one dimension to evaluate. Credibility generally incorporates criteria that can serve as cues to accuracy, but not equal to accuracy (e.g., triangulating evidence contributes to credibility, but does not require ground truth). Thus, this use of "accuracy"

by certain standards may further diversify interpretations of ratings, as well as the determinants considered during evaluation.

Another semantic issue concerns the rather liberal use of terms conveying certainty (e.g., confirmed). In intelligence contexts where the information "is always incomplete... [and] frequently ambiguous," (Tecuci et al.) these words could lead to excessive confidence on the part of the intelligence consumers. Compounding this issue is the tendency of analysts to confine their ratings to the higher ends of the scales, in their review of spot reports completed during a US Army field exercise, Baker, McKendry, and Mace (Baker et al.) highlighted that A1 and B2 represented 80% of all reliability/credibility ratings, with B2 alone comprising 74% of ratings. Allied intelligence doctrine explicitly discourages statements of certainty "given the nature of intelligence projecting forward in time" (NATO, 2016). However, it remains unverified whether "completely credible" actually conveys less certainty than confirmed by other sources. A piece of information may be confirmed by some sources and simultaneously disconfirmed by others.

2.1.2 – Source Reliability Determinants

To address miscommunication and misinterpretation originating from vague source history descriptors, this determinant could be quantified (e.g., source reliability = accurate information provided on total information provided). A quantitative method of tracking and updating source history could improve consistency and streamline the information evaluation process (Samet, 1975). However, this would fail to address the Admiralty Code's implicit treatment of source reliability as constant across different contexts (Capet and Revault D'Allones, 2014). Despite past performance, source reliability may vary dramatically depending on the type of information provided, characteristics of the source(s), and the circumstances of collection.

A Human Intelligence (HUMINT) source with a proven track record reporting on military operations may lack the expertise to reliably observe and report on economic developments. Beyond variable expertise, HUMINT source motivations, expectations, sensitivity, and recall ability may shift between situations, with major implications for information quality (Schum, 1987 and Pechan, 1995). Even the reliability of an 'objective source' (i.e., a sensor) is highly context dependent (Cholvy and Nimier, 2003). For example, inclement weather may compromise the quality of information provided by an optical sensor, despite a history of perfect reporting under ideal conditions.

Regardless of source history, most of the standards examined highlight reliability determinants such as “authenticity,” “competency,” and “trustworthiness.” The adoption of these determinants is consistent with the literature on source reliability (Schum, 1987; Cholvy and Nimier, 2003). However, the standards fail to define or operationalize these concepts. Their inclusion is therefore likely to increase subjectivity and further undermine the accuracy of reliability assessments. The standards examined also fail to operationalize the qualifiers used to describe each level. For instance, reliability ratings often incorporate whether an evaluator has “minor doubt,” “doubt,” or “significant doubt” about the source’s authenticity.

Aside from being vague, the use of modifiers (“minor,” “significant”) for some levels, and the unmodified term (“doubt”) for another, is critical because the unmodified term effectively subsumes the modified cases. Chang et al. [20] describe how a process designed to decompose and evaluate components of a problem (i.e., information characteristics) may exacerbate unreliability in assessments if that process is ambiguous and open to subjective interpretations. Given the ambiguity built into current standards, users are unlikely to retrieve every relevant determinant, let alone reliably and validly weigh every relevant determinant when arriving at an ordinal assessment.

Another issue with current source reliability standards is their failure to delineate procedures for evaluating “subjective sources” vs. “objective sources” (e.g., human sources vs. sensors) (Rogova, 2016), or primary sources vs. secondary/relaying sources (Lemercier, 2014) Source motivation may be relevant when assessing HUMINT sources, but not sensors. Similarly, source expertise may be highly relevant for a primary source collecting technical information (e.g., a HUMINT asset gathering information on Iranian nuclear technology), but not so much so for an intermediary delivering this information to a collector.

In cases where information passes through multiple sources, there are often several intervals where source reliability considerations are relevant (Lemercier, 2014). For instance, when receiving second-hand information from a HUMINT source, one might consider the reliability of the primary source, the reliability of the secondary/relaying source(s), the reliability of the collector, as well as the reliability of any medium(s) used to transmit the information (Lemercier, 2014).

Following initial collection, Nobel (Noble, 2009) describes how information may undergo distortion at other stages of the intelligence process. Just like sources, intelligence practitioners will vary in terms of their ability to reliably assess and relay information. For instance, an economic subject matter expert may lack the expertise to accurately evaluate and transmit information on enemy

troop movements. Beyond expertise, an intelligence practitioner's assessment is also undoubtedly influenced by his/her personal characteristics (e.g., motivation, expectations, biases, recall ability) as well as various contextual factors (Capet and Revault D'Allones, 2014; Schum, 1987 and Noble, 2009).

When a finished intelligence product is edited and approved for dissemination, managers may inject additional distortion by adjusting analytic conclusions (Noble, 2009). The many opportunities for distortion may warrant the formalization of information evaluation as an ongoing requirement throughout the intelligence process (Capet and Revault D'Allones, 2014). At the very least, efforts should be made to ensure intelligence practitioners and consumers are cognizant of the mutability of information characteristics following the initial evaluation.

2.1.3 – Information Credibility Determinants

Much like the source reliability standards examined, most of the information credibility scales suffer from an inherent lack of clarity. Information credibility generally includes confirmation "by other independent sources" as a key determinant. However, no guidance is provided as to how many independent sources must provide confirmation for that information to be deemed credible. Where one analyst considers confirmation by two sources sufficient for a confirmed rating, another might seek verification by three or more. Perceptions of how much corroboration is necessary may also vary depending on the information in question.

For instance, an analyst may decide that a particularly consequential piece of information requires more corroboration than usual to be rated confirmed. This lack of consistency could lead analysts to misinterpret each other's credibility ratings and consider pieces of information more or less credible than intended. The information credibility standards examined also lack instructions for grading pieces of information that are simultaneously confirmed and disconfirmed. Under the Admiralty Code, such information could be considered both confirmed / completely credible ('1') and improbable ('5') (Cholvy and Nimier, 2003).

Without guidance, an analyst may establish their evaluation more heavily on confirmed information, while others focus on disconfirmed information, or try to pursue a balance between confirmed and disconfirmed. These three approaches could generate very different evaluations, despite evaluating the same information. Capet and Revault d'Allonnes [6] argue that confirmation does not, in itself,

translate into information credibility, and that not all forms of confirmation should be weighted equally.

Theoretically, a spurious rumor corroborated by many unreliable sources (e.g., tweets about a second shooter during a terrorist attack), and disconfirmed by a single reliable source (e.g., a police statement indicating a single attacker), could still be rated highly credible under current standards. Capet and Revault d'Allonnes [6] advocate identifying a threshold whereby information must be confirmed by a clear majority, and undermined by few or no sources, while accounting for source reliability. This would directly contravene the Admiralty Code's treatment of source reliability and information credibility as independent.

Lesot, Pichon, and Delavallade (Lesot et al., 2014) note that current standards lack consideration of whether relationships of affinity, hostility, or independence exist between corroborating sources. Corroboration from a source that has a "friendly" relationship with the source under scrutiny should likely have less influence than corroboration from an independent or hostile source. If Saudi Arabia corroborates information provided by Syria (with which it has a hostile relationship), that confirmation should carry more weight than identical confirmation provided by Russia (which has a relationship of affinity with Syria).

Friendly sources should be expected to corroborate each other (Lesot et al., 2014). Friedman and Zeckhauser (Friedman and Zeckhauser, 2012) suggest that the current emphasis on consistency with existing evidence may encourage confirmation bias. "Biased attrition" is used to describe an information filtering process that systematically favors certain information types in a problematic way. Information that conflicts with prior beliefs and analysis may in fact be more valuable, as it can shift the views of analysts and consumers more significantly.

Friedman and Zeckhauser (Friedman and Zeckhauser, 2012) argue that credibility standards could reduce biased attrition by incorporating the extent to which information provides a new or original perspective on the intelligence requirement at hand. Capet and Revault D'Allonnes (Capet and Revault D'Allones, 2014) also suggest that current standards be modified to gauge the extent to which information provides "meaningful" corroboration. Along similar lines, Lemercier [22] notes that confirmation-based credibility standards do not account for the phenomenon of amplification, whereby analysts come to believe closely correlated sources are independently verifying a piece of information. In order to control for amplification, credibility evaluation could incorporate successive

corroboration by the same source, corroboration by sources of the same type, as well as comparative corroboration from different collection disciplines (Lemercier, 2014).

The current emphasis placed on confirmation/consistency may also reinforce order effects, given that new information must conform to prior information to be deemed credible. All else being equal, if an analyst receives three new pieces of information, the first item received will typically face the fewest hurdles to being assessed as credible. Meanwhile, the second piece of information must conform to the first, and the third must conform to both the first and second. Under this system, an analyst may inadvertently underweight information that is in fact more accurate or consequential than information received earlier, potentially decreasing the quality of analysis.

One option for dealing with order effects would be the formal inclusion of mechanisms to re-evaluate prior pieces of information as new information becomes available. Two of the US standards examined (US Dept. of the Army, 2010b and US Dept. of the Army, 2012b) advocate continuous analysis and re-evaluation of source reliability / information credibility as new information becomes available. However, neither document outlines a specific method for revaluation. Beyond confirmation, most of the information credibility scales examined incorporate consideration of whether an item is "logical in itself."

Current standards do not specify whether this simply refers to the extent that information conforms to the analyst's current assessment. Furthermore, the use of "not illogical" as a level between "logical in itself" and "illogical in itself" is nonsensical, as "not illogical" effectively means "logical" (in itself). As noted with regards to source reliability, the Admiralty Code's one-size-fits-all approach to information credibility neglects important contextual considerations. Several US standards suggest that credibility determinants have more relevance depending on the collection discipline(s) utilized. For example, TC 2-91.8 (US Dept. of the Army, 2010a) and ATP 2-22.9 (US Dept. of the Army, 2012a) suggest that there is a greater risk of deception (an information credibility determinant) when utilizing Open-Source Intelligence (OSINT) than Captured Enemy Documents (CEDs). Similarly, ATTP 2-91.5 (US Dept. of the Army, 2010b) refers to the Admiralty Code as the "HUMINT system," and recommends the development of separate rating systems to assess the three basic components of document and media exploitation (Document Exploitation [DOMEX], Media Exploitation [MEDEX], Cellphone Exploitation [CELLEX]).

Joseph and Corkill (Joseph and Corkill, 2011) stress that the Admiralty Code is a grading system rather than an evaluation methodology. Beyond what is outlined in the scales, evaluators may have

a formal assessment procedure and/or a more exhaustive list of determinants to consider. Supplementary documents add some clarity to the standards examined, but also vary in terms of which determinants are identified and emphasized. Additionally, none of these extra determinants are defined or operationalized, and may further contribute to subjectivity.

2.2 – Lack of independence between scores and multiple dimensions in a single score

One of the first examples of critiques towards the NATO was formulated by Baker, McKendry and Mace (Baker et al., 1968), three members of the US Army Behavioral Science Research Laboratory in 1968, as the NATO comprehensive review mentioned.

Over 1400 field intelligence reports of the US Army were analyzed, and a strong correlation between the source rating and the information credibility rating was found (87% of the reports felt along the diagonal A1, B2, C3, etc.), implying that the two dimensions were dependent, violating the most important aspect of breaking down intel in the two dimensions, as we explained earlier.

The potential for evaluators to inadvertently allow their judgment of the credibility of information to be influenced by the reliability of the source is a concern that should be considered when using the Admiralty Code. This is because the Code is designed to assess the credibility of information independently of the reliability of the source. However, the study findings suggest that this may not always be the case, even though we can't know for certain whether the objects of the study (i.e., army field intelligence reports) influenced the findings.

Baker et al. posit also that the multidimensionality of the information credibility scale is a potential source of error in its use. While source reliability is unidimensional, the information credibility is multidimensional, using a) consistency with other information and b) plausibility.

The use of multiple dimensions in a single scale can make it difficult to assess the credibility of information accurately, as evaluators may be influenced by one dimension (e.g., consistency with other information) more than another (e.g., plausibility). This is because the two dimensions are not necessarily correlated, meaning that information that is consistent with other information may not be plausible, and vice versa.

Additionally, the relative weighting of the two dimensions is not always clear, which can further complicate the assessment process.

To this point made by Baker et al., it must be though noted that even the source reliability is multidimensional. In fact, looking back at the inception of the NATO code, source reliability should be influenced by a) trustworthiness and b) competence.

2.3 – Readability and limited criteria

Besombes and Revault d'Alonnes (Besombes and Revault D'Allones, 2008) offer another set of criticisms of the NATO information evaluation system. They base their criticisms on STANAG 2022, which is the NATO standard for intelligence reports (containing the same tables as the AJP2).

The two French authors argue that the use of two axes in the NATO code makes it difficult to understand the communicated value of a score. They point out that it is not always clear whether information rated B3 or C2 is more probable. However, it must be noted that the rating system was never designed for exact comparison between pieces of information.

In this regard, we could perhaps quote McLachlan (McLachlan, 1939-45), who cautions against placing too much emphasis on the notion of a hierarchy of information: *"…. No piece of information is normally of great value on its own. When first received it is like a sentence without its context. Signal intelligence for example, in its raw state is seldom intelligible on its own to anyone but the expert who extracts it or deciphers it… it cannot be read and understood, even when translated, in isolation."*

Besombes and Revault d'Alonnes argue that the reliability of a source is independent of the information that the source provides. They contend that this means that all information provided by a reliable source should be considered equally reliable.

However, this is a flawed assumption. As McLachlan has argued in detail, the proficiency of the source should also be considered when assessing the reliability of the information. For example, your neighbor might be a very reliable source on barbecuing, but he's likely to be less reliable on COVID-19 vaccinations, unless he has a PhD in immunology.

The second point the authors make is that the NATO information evaluation system's criteria for determining the plausibility of information are too limited. They contend that the system should

also consider the proficiency of the source and the likelihood of the information. They propose that these additional criteria would be useful for expressing the confidence that an information deserves.

As already mentioned, proficiency is already reflected in the reliability score of the source. Separating this element from reliability in a distinct score might not add any value.

As for the additional element of likelihood, Besombes and Revault d'Alonnes (Besombes and Revault D'Allones, 2008) define it as a criterion that qualifies information based on our global understanding of the state of the world. However, this element could increase the risk of confirmation bias.

The authors finally propose to use a scoring chain to arrive at a confidence indicator that expresses, in a single digit, a combination of all four criteria. However, this approach is unlikely to lead to better information evaluations for two reasons:

First, the proposed scoring chain is relatively complex and may not be feasible for human-sourced intelligence (HUMINT) or open-source intelligence (OSINT). A complex scoring system might make more sense when speaking of sensory data, where evaluation is automated.

Second, condensing the final score into a single digit significantly reduces the communication value of the rating. A single digit score may not be sufficient to convey the full range of information about the reliability, plausibility, and likelihood of the information. This could lead to weak signals being overlooked.

2.4 – Communication and criteria

Another early scientifically rigorous critiques of the NATO Admiralty System, cited by the NATO comprehensive review, can be traced back to 1975 when Michael G. Samet, a researcher for the US Army's Research Institute for Behavioral and Social Sciences, conducted a series of experiments using the NATO Admiralty System (Samet, 1975).

The study, "Subjective Interpretation of Reliability And Accuracy Scales For Evaluating Military Intelligence", asked about 60 US army captains familiar with the grading system to evaluate 100 comparative statements. The findings were:

"Findings of the present study indicate that the two-dimensional evaluation should be replaced because:

1. The accuracy rating dominates the interpretation of a joint accuracy and reliability rating and

2. There is frequently an undeniable correlation between the two scales." (See 2.2 on lack of independence between the two dimensions by Baker et al.)

Irwin and Mandel, more recently (2019), offered a comprehensive critique of the application of the Admiralty Code (Irwin and Mandel, 2019). They argue that information evaluation methods mask, rather than effectively guide, subjectivity in intelligence assessment. Their critique of the NATO evaluation system consists of three parts:

Communication: The way that ratings are communicated can be misleading. For example, a rating of "A3" may be interpreted as meaning that the information is highly reliable, when it only means that the information is likely to be true.

Criteria: The rating determinants used in the NATO evaluation system are too limited. They do not consider the full range of factors that can affect the credibility of information, such as the source's motivation and the context in which the information was collected.

Structure: The position of information evaluation within the intelligence process is flawed. Information evaluators are often not given enough time or resources to do their job properly. They are also often not involved in the early stages of the intelligence process, when the most important decisions are made. We will not analyze this, as the organizational structure of military intelligence is not within our scope of research.

2.4.1 – Communication

Irwin and Mandel argue that the communicative value of the ratings in the NATO evaluation system is limited due to the following factors:

- **Subjective interpretations of the ratings:** the descriptions of the ratings in the Admiralty Code are not always clear, and different users may interpret them differently. This can lead to miscommunication between users, as they may not be interpreting the ratings in the same way.
- **Inconsistent ratings:** the ratings in the Admiralty Code are not always consistent with each other. For example, a source that is rated as "reliable" in one method may be rated as

"usually reliable" in another method. This can also lead to miscommunication, as users may not be aware of the different ways that the ratings can be interpreted.

- **Non-numerical values:** The descriptions of the ratings do not come with numeric values, which makes it difficult to compare ratings across different sources.
- **Use of terms that convey certainty:** such as "confirmed," can lead to overconfidence. This is because these terms suggest that the information is more certain than it is. This can lead to decision-makers making poor decisions based on the information.
- **Evaluators tend to confine their ratings to the high end of the scale:** such as "A1" or "B2." This is because they may be reluctant to give lower ratings, as they may be seen as being overly critical. This can lead to underestimation of the uncertainty of information, and to decision-makers making poor decisions based on the information.

These factors can contribute to miscommunication and misinterpretation of the ratings, which can have negative consequences for the intelligence process.

2.4.2 – Criteria

Irwin and Mandel contend that the rating determinants used in current evaluation methods are flawed because they do not account for situational considerations.

They argue that the Admiralty Code implicitly treats source reliability as constant across different contexts, which is a particularly problematic feature. While it is true that source reliability may be treated as constant in practice, this would be a flaw in the execution, not in the original idea behind the Admiralty Code. As described by McLachlan, the Admiralty Code does take different contexts into account.

Irwin and Mandel argue that most of the methods they examined highlight reliability determinants such as authenticity, competency, and trustworthiness. However, they point out that these methods fail to formally define or operationalize these concepts, which is likely to increase subjectivity and undermine the internal consistency of source reliability evaluations. Additionally, they point to the failure to distinguish between subjective and objective sources, such as human sources and sensors, or primary and secondary sources.

The Admiralty Code was originally devised to be applied in the context of human intelligence (HUMINT), so it is understandable that it does not distinguish between subjective and objective

sources. However, this is a valid point, as the distinction between primary and secondary sources is a recurring theme in the literature on information evaluation.

Here are some of the specific criticisms that Irwin and Mandel make of the rating determinants used in current evaluation methods:

- Not formally defined or operationalized.
- Subjective and therefore likely to vary among evaluators.
- Do not consider the context in which the information was collected.
- Do not distinguish between subjective and objective sources.
- Do not distinguish between primary and secondary sources.

These criticisms are valid and suggest that there is a need for further research on the development of more rigorous and objective rating determinants for information evaluation.

Irwin and Mandel make a valid point that information credibility is generally considered to be confirmed by other independent sources. However, they raise several important questions about this criterion, such as:

- How many independent sources must provide confirmation for information to be judged credible?
- Should relationships of affinity, hostility, or independence between the sources be considered?
- Does an emphasis on consistency with existing evidence encourage confirmation bias?

These questions suggest that the criteria of confirmation by independent sources is not as straightforward as it may seem. There is a delicate balance between confirmation and independence, and the specific criteria used to assess credibility will need to be carefully considered.

Irwin and Mandel's critique of the criteria used to assess credibility is insightful and thought-provoking. It highlights the need for further research on this topic.

2.5 – PSD2 – Revised Guidelines for Major Incidents Reporting

The European Banking Authority (EBA) revised Guidelines on major incident reporting under the Payment Services Directive (PSD2 were published in March 2021, providing additional and more detailed criteria on when an incident should be classified as major (EBA, 2021).

The updated guidelines define a major incident as an operational or security incident having, or likely to have, a significant impact on the provision of payment services.

Specifically, in order to be classified as "major" an incident must meet at least one "higher impact level" criteria or at least three "lower impact level" criteria. (Criteria listed afterwards).

The following criteria and underlying indicators must be evaluated by the payment service provider to assess the operational or security incident:

i. Transactions affected

The total value of the transactions affected, the number of payments compromised as a percentage of the routine level of transactions carried out with the services in question.

ii. Payment services users affected

The number of payment service users affected should be determined both in absolute terms and as a percentage of the total number of payment services users.

iii. Breach of security of network of information systems

Payment services providers should determine if any malicious action has compromised the security of the network or information systems related to the provision of payment services.

iv. Service downtime

The period of time during which the service will likely be unavailable for the payment service users, or during which the payment order cannot be fulfilled by the provider.

v. Economic impact

The monetary costs associated with the incident holistically and take into account both the absolute figure and the relative importance of these costs in relation to the size of the payment service provider.

vi. High level of internal escalation

Whether the incident has been or will likely be reported to their executive officers

vii. Other payment service providers or relevant infrastructures potentially affected

The systemic implications the incident will likely have, the potential spill over beyond the initially affected payment service provider to other providers, financial markets or payment schemes.

viii. Reputational impact

How can the incident undermine users' trust in the payment service provider affected, and more generally in the underlying service or the market as a whole.

Further general details are then given on the methodology of how to frame each indicator, what to consider and what not.

If no actual data is available to the service payment provider, estimates must be used in order to assess whether or not the threshold of the indicator has been reached, or it is likely that it will be reached before the incident is resolved.

Payment service providers should carry out this assessment on a continuous basis during the lifetime of the incident, so as to identify any possible status change, either upwards (from non-major to major) or downwards (from major to non-major).

All of these criteria must be then evaluated, in terms of the value of each and every indicator, in order to classify an incident. The classification is carried out by checking whether each criteria indicator falls in the minor or major incident threshold, and then, recalling what we said earlier, if at least one criterion is higher impact level, or at least three criteria are lower impact level, the incident is deemed as major.

The classification table is the following:

Table 1 - PSD2 major incident classification

Criteria	Lower impact level	Higher impact level
Transactions affected	> 10% of the payment service provider's regular level of transactions (in terms of number of transactions) **AND** duration of the incident > 1 hour **OR** > EUR 500.000 **AND** duration of the incident > 1 hour	> 25% of the payment service provider's regular level of transactions (in terms of number of transaction) **OR** > EUR 15.000.000
Payment service users affected	> 5.000 **AND** duration of the incident > 1 hour **OR** > 10% of the payment service provider's payment service users **AND** duration of the incident > 1 hour	> 50.000 **OR** > 25% of the payment service provider's payment service users
Service downtime	> 2 hours	Not applicable
Breach of security of network or information systems	Yes	Not applicable
Economic impact	Not applicable	> Max (0.1% Tier-1 capital, EUR 200.000) **OR** > EUR 5.000.000

Criteria	Lower impact level	Higher impact level
High level of internal escalation	Yes	Yes, and a crisis mode (or equivalent) is likely to be triggered
Other payment service providers or relevant infrastructures potentially affected	Yes	Not applicable
Reputational impact	Yes	Not applicable

The Guidelines then continue with the explanation of the notification process to the competent authorities, which is not within the scope of our dissertation.

The PSD2 directive updated Guidelines are not intended for a-priori, potential-threat-evaluation analysis. They are intended for ex-post incident reporting, consisting of three reports, to be filed in three separate phases of the incident lifetime.

The initial report must be filed as soon as the incident has been classified as major, the intermediate report when regular activities have been restored and business is back to normal, and the final report when the root cause analysis has been completed.

Thus, in order to use the parameters and principles enunciated in the Guidelines for a a-priori threat evaluation system, some generalization must be made, and a lower level of detail is required in order to fulfill the purpose of an a-priori threat evaluation system, the immediate readiness to support security decision-making, in this sense, caution will be more important than precision.

3 – A new and expanded intelligence grading system for the financial sector

We now have every ingredient at our disposal to create a new threat-intelligence grading system, specific to the financial sector.

We know which the major and more influential critiques are to the current intelligence grading system, the NATO Code, thus, how we could advance it.

We know which parameters are currently used to classify incidents under the PSD2 directive, thus providing a possible structure for a threat-intelligence classification system in the field of financial services.

By combining these two elements, we can then craft a new and expanded threat intelligence grading system for the financial sector, capable not only of evaluating source credibility and information plausibility, but also the potential impact on the victim of the threat and the potential systemic diffusion the threat might encompass, as we laid out in the first chapter.

The system we will propose is of course just a suggestion, a starting point, intended as a steppingstone in order to perhaps ideate more complex or more detailed grading systems for threat-intelligence, specializing in the field of reference, or maybe even capable of being generalized.

3.1 - Methodology

After having carefully reviewed the literature and further discussions with field experts, we convened that the system should take into consideration as few parameters as possible, in order to not overcomplicate the grading process and reduce the whole process time. The parameters should thus be able to be as widely applicable as possible to the different entities affected by the possible threat. At the same time, they must be as pregnant as possible in terms of being able to capture the real level of danger the threat poses.

The choice of the specific parameters we will take into consideration to formulate and propose our new grading model, are the result of conversations and discussions with mainly two “sets” of experts: cyber threat analysts from CERTFin and officers from the naval intelligence divisions of the

Italian Navy. I will now try to summarize what their contributions were, as they were the main drivers of inspiration behind the choice of parameters and scope of the work overall.

3.1.1 – Contributions: CERTFin Cyber Threat Analysts

CERTFin is a private and public cooperative initiative, funded by the Italian Banking Association (ABI) and the Italian Central Bank. Its mission is that of augmenting financial actors cyber risk management, increasing the cyber resilience of the whole Italian financial sector. Providing operational and strategic support to prevention and response activities to cyber attacks and security incidents. Finally, it helps national and international cooperation by facilitating information sharing. I met the people working at CERTFin thanks to my Professor and relator, Paolo Spagnoletti, who put me in contact with an executive.

The meetings were immediately fruitful, discussing with the cyber threat analysts, the idea of expanding the current threat evaluation system arose, mainly because they considered necessary to have a formal process and grading system to classify the threats they encounter and more importantly they deemed critical to have a widely agreed way of communicating the threat level. They narrated various examples of Italian banks and agencies going into panic mode just because the cyber threat level they were facing was not communicated appropriately by the authorities involved. This, in their view, was deemed as a serious and impellent issue.

CERTFin and other cyber threat analysts worldwide already make extensive use of the NATO Code to estimate source reliability and information credibility, but they thought that, being in a field evaluating only threats, the possibility of expansion and inclusion of an impact and systemic diffusion scale really interested them. The PSD2 Guidelines as inspiration for the financial part of the new grading framework came from one of their suggestions, as they had analyzed the guidelines in the past, in order to classify incidents as either “major” or “minor”, and we agreed that the parameters presented in the Guidelines could have been fitting and appropriate for the grading system.

3.1.2 – Contributions: Italian Navy Intelligence Officers

As mentioned in the Foreword, a big inspiration and later support for writing this thesis were in fact the many exchanges of views I had during my time in the Italian Navy Mare Aperto EX with several

Naval Intelligence Officers. While of course the financial sector is not what they are familiar with, they were extremely precious when they explained the actual usage and process behind the evaluation of intelligence. As much as one can read and study from the literature, a relatively distant field from an everyday normal life, as the intelligence world is, is not really fully graspable until one comes into contact with it. The processes, human relations, technical hurdles and conscience doubts that intelligence officers experience is something that is difficult to perceive from a study or research paper, or even from books or movies.

While not of course very familiar with the financial world, all of the Naval Intelligence Officers I exchanged views with, allowed me to gain enormous and invaluable insights on the path to take to develop the new framework, by sharing the ideas I had developed together with the CERTFin analysts with people who handle all kinds of intelligence (not just threat intelligence) everyday.

The main points I took away from their views and opinions were that, while complicated to fabricate, in terms of producing a scale that the majority of field experts would agree on, the idea of expanding the system to evaluate also impact and systemic diffusion was very clever, they even mentioned that sometimes they try to "forecast" the potential impact, when the intelligence is about a threat, in order to put the consumer of that intelligence in a better position to make decisions. Another point I took away from them is that, while they all recognized the shortcomings and fallacies of the NATO Code (see Chapter 2), it is so widely used and crystallized in everyday intelligence life, that it would probably have been a case of "trying to reinvent the wheel" if I were to change it all. But they thought those same shortcomings and fallacies, being so well-known, would have been exceptionally valuable to construct the other part of my grading system. Which is in fact exactly what I've tried to do, refining and polishing the aspects where the NATO Code falls short (see 3.3 further on).

For these reasons, while maintaining intact the first part from the NATO Admiralty System, the parameters we chose to expand the system, extracted from the PSD2 directive updated Guidelines, are:

Potential impact:

- Transactions affected
- Users affected

Potential Systemic diffusion:

- Other payment service providers or relevant infrastructures potentially affected
- Reputational impact

3.2 - Rationale and grading tables

The NATO grading system works on a scale from 1 to 6, as we've seen in the previous chapters, while the PSD2 directive parameters works on major/minor incident level, thus, we chose to adapt the guidelines parameters in order to fit on the 1 to 6 scale. .

The levels we propose are the result of careful analysis of experts' opinions.

Below are the resulting grading tables:

Table 2 - Potential impact dimension suggested grading table

Potential impact		
Level/Criteria	**Transactions affected**	**Users Affected**
a	**50% < x < 100%** of the payment service provider's regular level of transactions (in terms of number of transactions) OR > EUR 30.000.000	**50% < x < 100%** of the payment service provider's payment service users OR > 100.000
B	**25% < x < 50%** of the payment service provider's regular level of transactions (in terms of number of transactions) OR > EUR 15.000.000	**25% < x < 50%** of the payment service provider's payment service users OR > 50.000
C	**10% < x < 25%** of the payment service provider's regular level of transactions (in terms of number of transactions) OR > EUR 2.000.000	**10% < x < 25%** of the payment service provider's payment service users OR > 15.000

Potential impact		
Level/Criteria	**Transactions affected**	**Users Affected**
D	**5% < x < 10%** of the payment service provider's regular level of transactions (in terms of number of transactions) OR > EUR 500.000	**5% < x < 10%** of the payment service provider's payment service users OR > 5.000
E	**1% < x < 5%** of the payment service provider's regular level of transactions (in terms of number of transactions) OR > EUR 100.000	**1% < x < 5%** of the payment service provider's payment service users OR > 500
F	**0% < x < 1%** of the payment service provider's regular level of transactions (in terms of number of transactions) OR > EUR 0	**0% < x < 1%** of the payment service provider's payment service users OR > 0

Table 3 - Systemic Diffusion suggested grading table

Level/Criteria	Other payment service providers or relevant infrastructures potentially affected (in the country of reference)	Reputational impact
1	**> 50%** of payment service providers affected	Wide media coverage, high victim reputational damage, **high reputational damage to the whole system**
2	**> 25%** of payment service providers affected	Wide media coverage, high victim reputational damage, **some reputational damage to the whole system**
3	**> 5** other payment service providers affected **AND** **>= 1** relevant infrastructure affected	Incident reported on mass media, **some reputational damage with public** (limited to the victim)

Level/Criteria	Other payment service providers or relevant infrastructures potentially affected (in the country of reference)	Reputational impact
4	**>= 1** other payment service provider affected **AND** **>= 1** relevant infrastructure affected	Incident reported just on specialized media, **some reputational damage with field experts**
5	**= 1** other payment service provider affected	Incident reported just on specialized media, **no reputational damage with field experts**
6	The threat is **limited to the victim** (no one else uses the system the threat exposes)	Incident is **not reported in the media**, **no reputational damage**

Just as in the PSD2 Guidelines in absence of actual data the entity assessing the incident must act on estimates, in the case of threat-analysis all of these parameters must be estimated, as obviously a threat must be evaluated in terms of what it could cause.

Hopefully, with due caution, when unsure in which category the threat falls in, it's advisable to put it in the higher-level.

3.3 - Learning from literature critiques to NATO Code

The careful and thorough literature review we conducted on the critiques put forward towards the NATO Grading System, allowed us to sharpen some aspects of our own grading grids.

In particular, we found some of the points argued by Irwin and Mandell the most important, for this reason, we would like to focus on some details of the proposed grading system, explaining where the improvement, steering from the aforementioned critiques, lies.

- Three out of four variables to be considered to evaluate the threat level have **numerical values**, thus drastically reducing the margin for a subjective interpretation of a danger. For what concerns reputational damage, we found it to be more fitting by verbal description.
- Having numerical values, it's much more difficult for the grading of one variable to influence the grading of another, during the analysis process of evaluation, if it were to happen (for example, a positive correlation might be hypothesized between reputational damage and

number of users affected), it would probably not be a judgmental error, but the proof of an underlying correlation.

- By taking into consideration two separate dimensions, not concerning the same realm of interest (they can move independently), we find that this system is much more easily understandable in terms of communication, meaning that the letter and the number express two very different aspects of the threat, thus not creating confusion. For example, while an A3 or C1 grading for source and information credibility can be considered more or less equal, an a3 or c1 grading for impact and systemic diffusion are clearly not equal. The first would be completely impeding for the victim of the threat, resulting in a huge reputational damage, but leaving the public trust in the whole system virtually untouched. The second one would instead be moderately impeding for the victim, but so systemic that the whole system would suffer great loss of trust from the wide public.

3.4 - How should threat be evaluated

We thought it would make sense to keep the NATO taxonomy in place, in terms of the alphanumeric score, to not have to diverge from the current evaluation framework, but to hopefully seamlessly integrate with the current threat evaluation processes. Thus, reproducing the same taxonomy, letter for impact (non-capital) and number for systemic diffusion.

As you will have noticed, both evaluation dimensions (impact and systemic diffusion) have two parameters to be considered corresponding to a threat level.

This is not to make the grading more complex, but instead to make it easier, erring on the side of caution. By this we mean that, for example, if the potential impact of a threat is deemed level “c” in terms of number of transactions affected, and “d” in terms of users affected, we think it’s reasonable to take the highest of the two levels to grade the threat.

The rationale behind this choice is that, as we mentioned earlier, the fundamental purpose of a threat-intelligence grading system is that of supporting decision-making, thus the more cautious the prevision, the safer the outcome. As the common sense saying goes: “Hope for the best, prepare for the worst”. We deemed this approach fitting to the situation.

3.5 - What can be improved

The most immediate, and most reasonable, critique we can think of the grading system we just presented is, of course, that the levels are subjective, not precise in terms of reflecting a “step” in the danger levels, thus, creating a useless classification grid.

Another justified critique might be that the parameters chosen to evaluate both of the dimensions are not really fitting, or that they are too few, or too many.

Further studies and research might delve deeper into many aspects of a grading system like the one we are proposing here, with the aim of advancing it and improving it, for example:

a) Are impact and systemic diffusion really the two most important dimensions to estimate when evaluating threat-intelligence information?
b) Are the two variables chosen for potential impact the best possible ones?
c) Are the two variables chosen for systemic diffusion the best possible ones?
d) Are the levels chosen for each variable the best possible ones?

It is nonetheless our opinion that, while the abovementioned critiques do have grounds to base themselves on, this is a needed starting point for cyber threat intelligence.

Further studies and research might and should critique the system we just put forward, trying to find better parameters, variables, more fitting levels, more suitable descriptions, and so on.

4 – Evaluating the new system

For the design and evaluation process of the new threat-intelligence grading system we just proposed, we considered this work, identifiable as the design of a new evaluation methodology, as Design Science Research (DSR). The framework we used for the design and evaluation of the DSR project is that developed by John Venable et al. in the European Journal of Information Systems: "FEDS: a Framework for Evaluation in Design Science Research" (John Venable et al., 2016).

In their paper, the authors identify a research gap in the DSR process, by stating that no comprehensive and coherence guidance was at the time available, especially for novices just starting out in the field.

For this reason, Venable et al. sort of lay out a step-by-step guide to guide an initial project of DSR, in order to ensure the rigorousness, validity and relevance of the work. We will be following their FEDS to achieve precisely those same goals.

The FEDS framework unravels the DSR design and evaluation process in 4 proposed steps:

1. Explicate the goals
2. Choose the evaluation strategy
3. Determine the properties to evaluate
4. Design the individual evaluation episode(s)

We'll analyze each step and how we approached it in dedicated paragraphs, to validate our proposal and ensure rigorousness in work.

4.1 – Explicate the goals

The FEDS framework identifies four possibly competing goals in designing the evaluation component of Design Science Research.

Rigour: Intended in two senses:

1. Efficacy: establishing that it is the artefact instantiation causing an observed outcome, and only the artefact, not some confounding independent variable or a contingent situation.

In our case: not really applicable as, by proposing an evaluation method for threat-intelligence, the final outcome produced by the evaluation method would be the actions and countermeasures undertaken by the decision-makers in the threatened organization.
Thus, those actions would be subject to personal, contingent and specific situations that are not dependent on the evaluation method itself, but rather the efficacy could be measured only with prolonged and established use over time of the system, highlighting potential downfalls or evaluation needs not taken into account.

2. Effectiveness: establishing that the artefact instantiation works in a real situation.

   In our case: by expanding an already established and affirmed evaluation method for threat intelligence, we could hypothesize that, in terms of working, it already works. In the worst case scenario, the system could add some attrition in learning and applying the new dimensions for the evaluation, but we believe we can be fairly confident that it is reasonable to think that the system would work.

Uncertainty and risk reduction:

Risks can be human, social, use or technical. It is important to identify what the potential risks might be in advance, to influence and improve future design and development.

In our case: we have already identified the potential downfalls of the method, in particular the dimensions chosen for the evaluation might not be enough to fully represent the threat the organization is facing, or they might be stringent, and an organization may not find itself capable of assigning a threat level.

For future development of the system, we think it would be necessary to see it in action, used in a real threat-intelligence evaluation scenario, only at that time those critical points could manifest themselves.

Ethics:

The evaluation of an artefact should estimate potential risks to animals, people, organizations or the public.

In our case: we find reasonable to think that by providing additional elements to an already existing threat-intelligence evaluation method, no real additional risks could emerge. The risks associated with the method would probably be the same of the already existing NATO Admiralty System, mainly the over or under estimation of a threat, depending on personal or organizational interests, differing from that of genuinely assessing a possible threat to the organization itself, or worse, its users and the general public.

Efficiency:

Efficient evaluation aims at balancing the aforementioned goals with he resources available for the evaluation itself (time, money, people availability, etc.).

For this issue, it is necessary to introduce the concepts of formative vs naturalistic evaluation. While formative evaluation means evaluating the artefact characteristics, associated risks, possible usage problems, etc., *a-priori*, from a theoretical perspective, naturalistic evaluation means testing out the artifact in its "natural" environment from the start, to ensure the maximum possible degree of effective testing and feedback.

In our case: while a naturalistic evaluation is almost always the best possible choice in terms of completeness of information, it can be often too costly, in terms of money, resources and time needed for an appropriate evaluation. Especially in our case, to deploy the system in a real environment would have required extremely demanding approval processes (threat-intelligence evaluation in the financial sector is done mainly by government or para-government agencies), we thus judged it to be too costly. For this reason, we chose a formative evaluation. As you will see later on, we simulated some fictitious scenarios (expository instantiation) to evaluate the model.

4.2 – Choose a strategy for the evaluation

Choosing a strategy for evaluating means choosing when, for what purpose and how to evaluate, the FEDS framework proposes four alternative strategies:

- **Quick and simple:** little formative evaluation, quickly progressing to more naturalistic evaluations with few evaluation episodes. Low cost and rapid, not advisable in the presence of many and various design risks.

- **Human Risk and Effectiveness:** more emphasis on formative evaluations, quickly progressing to more naturalistic ones, rigorous evaluation of the artefact effectiveness in real environments and in the longer run, in spite of the humans and social issues associated with adoption and usage.
- **Technical Risk and Accuracy:** iterative formative evaluations, progressing towards summative artificial evaluations to determine efficacy of the artefact, naturalistic evaluations carried out at the end of the evaluation process.
- **Purely Technical Artefact:** when no human user is involved, that is, the artefact is solely technical, or when the deployment with users is so far in the future that naturalistic strategies would be considered irrelevant.

Based on the goals of the evaluation, one or more strategies could be appropriate. Each strategy one chooses to employ entails decisions on why, when and how to evaluate. When choosing the strategy for the evaluation, the following heuristics must be considered:

1. Evaluate and prioritize design risks, understood as major problems that the design may face.
2. Evaluate how costly it would be to evaluate the artefact with real users in the real setting.
3. Evaluate whether the artefact is purely technical.
4. Evaluate whether the construction of the design is small and simple or large and complex.

In our case: as we've already state, we judged the deployment into real settings with real users to be too costly. Moreover, we do not envision deployment of the model into real settings in the foreseeable future, as we believe that before deployment a thorough process of insiders' evaluation and feedback would be necessary.

For all the above reasons, we've identified the Purely Technical Artefact Strategy as the most fitting and suitable one for our work. This will imply solely theoretical evaluation and no real-life deployment.

4.3 – Determine the properties to evaluate

What should be evaluated? What general set of features, goals and requirements must be subject to evaluation?

Each artefact, within its situation, environment and contingent variables will have idiographic practical requirements. The selection of the properties to evaluate is necessarily unique to the artefact, revolving around its purpose, context and situation.

Based on the goals and strategy, a different set of unique properties will arise.

In our case: we considered several heuristics to identify which properties we should evaluate, briefly summarizing:

- Recalling the goals we set ourselves to achieve:
    1. Identify variables capable of estimating a threat impact on the potential victim
    2. Identify variables capable of estimating a threat systemic diffusion
    3. Merge these variables into a NATO-like threat intelligence nomenclature

- Recalling the critiques put forward by the literature to the NATO Admiralty System:
    1. No precise boundaries within levels (descriptive, non-numeric classification
    2. Risk of one evaluation dimension influencing the other (source and information)
    3. Simple communication and level understanding

We identified as pregnant evaluands the following questions:

- Are the variables chosen for the classification capable of estimating the threat impact and potential systemic diffusion?
- Is the new classification coherent with the NATO standard?
- Are the category levels clearly defined and separate?
- Are the evaluation dimensions independent from each other?
- Is the resulting output easily communicable and understandable?

4.4 – Design the individual evaluation episode(s)

The strategy is chosen and the properties to evaluate are determined.

The actual evaluation must now be designed, taking into consideration several elements, such as the constraints in available resources and in the environment, prioritizing essential elements, how many evaluation episodes and when will they be carried out.

In our case: not having access to real settings or users, the evaluation process, as stated before, will be purely theoretical, also considering the Purely Technical Strategy chosen. Thus, formative evaluation will be conducted in advance, and an expository instantiation ex-post will be conducted to evaluate the effectiveness of the artefact.

4.5 – Fictitious scenario (expository instantiation)

No publicly available information on current or past cyber threats to financial institutions or payment service providers are available. As we've already mentioned, these kinds of information are strictly confidential, as releasing information about an internal vulnerability would not only expose the organization to further exploitations, but also damage its reputation.

Therefore, to evaluate our proposed cyber threat-intelligence grading system, we will present a fictitious scenario, starting from a known and existing threat and ideating an ad-hoc payment service provider as the victim of the threat and possible consequent attack.

4.5.1 – The threat: Egregor

The cyber threat we will take into consideration is a well-known ransomware attack: Egregor.

Egregor is a sophisticated ransomware attack first discovered in September 2020, it is a variant of the Sekhmet ransomware, believed to be originated with the Maze hacker group. Egregor is distributed as a *Ransomware-as-a-Service (RaaS)* and the group who operates it adopts the double extorsion technique (CSIRT Italia, 2021).

First, it exfils the victim's data, then it encrypts that very same data. The group demands the payment of the ransom in order to both decrypt the original data and to not have that data published on the "Egregor News-Hall of shame", which occurs if the victim does not pay in due time.

CSIRT Italia, after having analyzed the victimology of Egregor, deducts that the main targets are private or public entities with global exposure, capable of paying a relevant ransom. Especially targeting tech, financial, health, government and manufacture sectors.

4.5.2 – Egregor Kill Chain

We will now summarize here the main phases of the Egregor Kill Chain:

- **Reconnaissance:** targeting Big Game Hunting.
- **Weaponization:** creation of infected documents or files (hidden DLL files).
- **Delivery:** sending emails with malevolent macro attachments.
- **Exploitation:** infected macro execution, that then download various malware families (Qbot, Urnisf, IcedID, etc.).
- **Installation:** the malware installation ensures exfil of valid access and privilege credentials needed for lateral movements, continuing reconnaissance and pursue persistence.
- **Command and Control (C2):** through the post-exploitation software "Cobalt Strike", a line of communication towards C2 is established, allowing further downloads of scripts, DLLs and other useful files.
- **Actions on objectives:** before encrypting, data is exfilled via FTP towards the attacker infrastructure using a version of Rclone, used to manage remote storage. Egregor then encrypts both local and shared files, keeping the OS functioning. In the end, notes for the ransom payment are generated.

4.5.3 – The victim: SecurePay

This is the fictitious part of our scenario. No real data has ever been publicly released by an organization on the technicalities or the specifics of a received attack, or potential vulnerabilities. Thus, we'll need to create a fictitious entity, of whom we know everything.

Let's take an imaginary payment service provider: "SecurePay", an emergent payment service provider serving millions of users worldwide, know for fast, efficient payment processing and user-friendly interface.

Let us also suppose SecurePay has a total user base of 10 million users, among them both private customers and various businesses. It manages around 1 million transactions per day (for reference, VISA processes more than 700 million transactions per day, as per the VISA Fact Sheet, FY23Q3 (VISA, 2023)).

Let's take the average transaction processed by SecurePay as a EUR 100 transaction (VISA is approximately around EUR 55).

4.5.4 – Threat detection

Now let's imagine that a third-party cybersecurity consulting firm, or the internal cybersecurity team in SecurePay stumbles upon a piece of information on one of the several Threat Intelligence Platforms they monitor daily. The information is an extract of a dispatch released by CSIRT Italia, an entity within ACN, the Italian National Agency for Cybersecurity. The dispatch mentions a new strain of the infamous Egregor ransomware, warning that organizations using the anti-malware system "CrowdSpersky" (fictitious) might be vulnerable as the anti-malware has a vulnerability that can be exploited in order to disable the automatic blocking of downloads from an external source.

One of the analysts, knowing that SecurePay uses exactly CrowdSpersky, is immediately alarmed and starts an evaluation of the threat in order to give the CISO more information to make decisions.

4.5.5 – Grading the threat

The analyst uses our new model, starting from the evaluation of source credibility and plausibility of the information. The source is an institution, specialized in cyber security matters, thus a level of "A" is given to source credibility. The information involves internal CrowdSpersky data, which, even though the source might have had access to, is probably not 100% accurate or up to date, thus the analyst assigns "2" to information plausibility. The first part of the score is thus evaluated as A2.

If the analyst were to stop here, the CISO would not have had any idea on whether this threat was actually a real problem to the SecurePay, how many people would have been needed to be assigned to preventive analysis and monitoring, how many resources should have been allocated to the threat.

Using our new model, the analyst instead estimates that, approximately 10 to 20% of the daily transactions processed by SecurePay could be compromised, relying on past internal data and analysis on incident response time. Also, 20 to 30% users could be affected, per internal data and analysis. The transactions affected variable would clearly fall in the "c" level boundaries (10 to 25% and the analyst estimates 10 to 20%). But the users affected variables goes from the "c" level (10 to

25%) to the “b” level (25 to 50%) as the estimates fall between 20 to 30%. As we’ve suggested earlier, when evaluating threats, caution is a virtue, thus the analyst will consider the “b” level.

The higher of the two variables for the potential impact is thus estimated as “b”, which is the grading given.

For the systemic diffusion variables, the analyst estimates that the CrowdSpersky anti-malware is used basically only by SecurePay, and however no other payment service provider or relevant infratrucure uses it, thus a score of “6” is assigned to the Other payment service providers or relevant infrastructures potentially affected.

The reputational damage is instead estimated higher, being SecurePay an up-and-coming company, often cited in the media, the analyst foresees high reputational damage with the public for SecurePay, and some system reputataiona damage, thus a score of “2” is assigned to the reputational damage variable.

The higher of the two variables is “2”, thus the score assigned to the Systemic Diffusion dimension is “2”.

Putting it all together, the two dimensions produce a grade of b2, thus making the threat posed by Egregor a high priority task in terms of impact and diffusion.

The outcome of A2b2, confirms the Egregor threat as a high priority, high risk threat that deserves the SecurePay cybersecurity team full attention.

4.6 - Discussion

We find reasonable to believe that an outcome of A2b2 would mean that the hypothetical SecurePay CISO diverts the team’s complete resources and attention towards trying to prevent and counteract the threat, either by changing the anti-malware partner, by working on developing a patch to the system that compartmentalizes the threat, or other possible actions.

Anyhow, the CISO would certainly have a firmer grip on the situation, and could prioritize time and resources with more efficacy, as the threat level would be clearly defined.

For what concerns the evaluation strategy we chose, the evaluands we had to consider can now find an appropriate answer:

- Are the variables chosen for the classification capable of estimating the threat impact and potential systemic diffusion? **Yes, if the data mentioned is available.**
  Is the new classification coherent with the NATO standard? **Yes, it maintains the 1 to 6 scale and nomenclature (alphanumeric)**
- Are the category levels clearly defined and separate? **Yes, the numeric boundaries clearly define each level.**
- Are the evaluation dimensions independent from each other? **Yes, not even variables influenced each other.**
- Is the resulting output easily communicable and understandable? **Yes, the A2b2 outcome clearly makes the threat high priority.**

Of course, the methodology used to evaluate the model has many flaws. First, as we repeatedly stated, not having any real data, and not being able to test the system in a real environment are huge handicaps for a proper, rigorous and valid evaluation of the model performance.

Future research and studies could certainly expand on the issue, by either testing the model in a real setting or by testing it with real data (maybe someone already working in the field).

These kinds of test would certainly help in identifying flaws, gaps and misconceptions implied in the model.

5 - Conclusions

The new and revised grading system for evaluating cyber threats in the financial sector we proposed hopes to be the first step in a much-needed direction, that of better understanding of the cyber threats the financial world is faced with, while helping analysts and decision-makers in the field to be more informed and conscious about what they are dealing with.

It also hopes to be a first step in the literature, as we've seen in fact, several and various critiques have been made in the last 50 years to the NATO Admiralty System, but a formal proposal, an expansion, is still yet to bee seen. While we do not of course envision this system to become the new standard (even more as it is limited to the financial sector), we surely hope it will spur discussions and research on the current state of threat intelligence, intelligence evaluation and communication, surely it marks an interesting contribution to the existing literature, by providing a possible case or example from which further research can move forward.

Many aspects of the grading system we proposed must be revised and improved, as we've tried to explain even in this paper. Future studies and research should focus on choosing more fitting parameters for the grading scales, and maybe identify more suitable boundaries between the levels. It could also be possible to instead try to generalize the system, thus abstracting it from the financial lenses it wears, and trying to make it for threat intelligence as a whole, independently of the field of application.

Future research could also study the possibility of re-evaluating the first part of the grading system, the standard NATO Admiralty System, by proposing a better, renovated and clearer system.

Bibliography

[1] - AJP-2 (Feb 2016) – NATO Standard – Allied Joint Doctrine for Intelligence, Counter-Intelligence and Security (Ed. A, Ver. 2).

[2] - NATO Osint Handbook V1.2 (2001).

[3] **–** Unver, A. (2018) Digital open source intelligence and international security: a primer. EDAM Research Reports, Cyber Governance and Digital Democracy 8.

[4] - Clusit (2022), Rapporto 2022 sulla sicurezza ICT in Italia.

[5] - Samet, M. G. (1975) 'Subjective Interpretation of Reliability And Accuracy Scales For Evaluating Military Intelligence', US Army's Research Institute for Behavioral and Social Sciences.

[6] - Capet, P., and Revault d'Allonnes, A. (2014). Information evaluation in the military domain: Doctrines, practices, and shortcomings. In: Information Evaluation, Capet, P., and Delavallade, T. (Eds.), 103-125. Hoboken, NJ: Wiley-ISTE.

[7] - United States Department of the Army. (2006). Field Manual FM 2-22.3, Human Intelligence Collector Operations. Washington DC.

[8] - United States Department of the Army. (2010a). Training Circular TC 2-91.8, Document and Media Exploitation. Washington DC.

[9] - Department of National Defence. (2011). Canadian Forces Joint Publication CFJP 2-0, Intelligence. Ottawa, ON.

[10] - United States Department of the Army. (2012a). Army Techniques Publication ATP 2-22.9, Open-Source Intelligence. Washington DC.

[11] - United States Department of the Army. (2012b). Army Techniques Publication ATP 3-39.20 Police Intelligence Operations. Washington DC.

[12] - United Kingdom Ministry of Defence. (2011). Joint Doctrine Publication JDP 2-00, Understanding and Intelligence Support to Joint Operations, (3rd ed.) Swindon, UK.

[13] - United States Department of the Army. (2010b). Document and Media Exploitation Tactics, Techniques, and Procedures ATTP 2-91.5 – Final Draft. Washington DC.

[14] - Tecuci, G., Boicu, M., Schum, D., and Marcur, D. (2010). Coping with the Complexity of Intelligence Analysis: Cognitive Assistants for Evidence-Based Reasoning. Research Report #7, Learning Agents Center. Fairfax, VA: George Mason University.

[15] - Baker, J.D., McKendry, J.M., and Mace, D.J. (1968). Certitude Judgements in an Operational Environment. Technical Research Note 200. Arlington, VA: US Army Research Institute for Behavioral and Social Sciences.

[16] - Samet, M.G. (1975). Quantitative interpretation of two qualitative scales used to rate military intelligence. Human Factors 17 (2):192-202.

[17] - Schum, D.A. (1987). Evidence and Inference for the Intelligence Analyst. Lanham, MD: University Press of America.

[18] - Pechan, B.L. (1995). The collector's role in evaluation. In: Inside CIA's Private World: Declassified Articles from the Agency's Internal Journal, Westerfield, H.B. (Ed.), 99-107. New Haven, CT: Yale University Press.

[19] - Cholvy, L., and Nimier, V. (2003). Information evaluation: Discussion about STANAG 2022 recommendations. In: Proceedings of the NATO-IST Symposium on Military Data and Information Fusion. Prague, Czech Republic.

[20] - Chang, W., Berdini, E., Mandel, D.R. and Tetlock, P.E. (2018). Restructuring structured analytic techniques in intelligence. Intelligence and National Security 33 (3):337-356.

[21] - Rogova, G.L. (2016). Information quality in information fusion and decision making with applications to crisis management. In: Fusion Methodologies in Crisis Management, Higher Level Fusion and Decision Making, Rogova, G.L., and Scott, P. (Eds.), 65-86. Cham, Switzerland: Springer International Publishing.

[22] - Lemercier, P. (2014). The fundamentals of intelligence. In: Information Evaluation, Capet, P., and Delavallade, T. (Eds.), 55-100. Hoboken, NJ: Wiley-ISTE.

[23] - Noble, G.P., Jr. (2009). Diagnosing distortion in source reporting: Lessons for HUMINT reliability from other fields. Master's thesis. Erie, PA: Mercyhurst College.

[24] - Lesot, M., Pichon, F., and Delavallade, T. (2014). Quantitative information evaluation: Modeling and experimental evaluation. In: Information Evaluation, Capet, P., and Delavallade, T. (Eds.), 187-228. Hoboken, NJ: Wiley-ISTE.

[25] - Friedman, J.A., and Zeckhauser, R. (2012). Assessing uncertainty in intelligence. Intelligence and National Security 27 (6):824-847.

[26] - Joseph, J., and Corkill, J. (2011). Information evaluation: How one group of intelligence analysts go about the task. In: Fourth Australian Security and Intelligence Conference. Perth, Australia.

[27] - Besombes, J., and A. Revault d'Alonnes (2008) An Extension of STANAG2022 for Information Scoring. In 'Proceedings of the 11th International Conference of Information Fusion' p. 1635-1641.

[28] - Room 39: Naval Intelligence in Action, 1939–45. By McLachlan Donald. London: Weidenfeld & Nicolson.

[29] - Irwin, D. and D. Mandel (2019) 'Improving information evaluation for intelligence production', Intelligence and National Security, Vol. 34(4): pp. 503-525.

[30] – European Banking Authority (2021), revised Guidelines on major incident reporting under the Payment Services Directive (PSD2).

[31] - John Venable et al. (2016) - FEDS: a Framework for Evaluation in Design Science Research; European Journal of Information Systems.

[32] - EGREGOR: Attività di un ransomware; CSIRT Italia, Maggio 2021.

[33] - VISA Fact Sheet, FY23Q3.