FÉVRIER 2022

# WHITE PAPER

## CONCEPT CODE: SOLUTION AGAINST CYBER ATTACKS?

GLIMPS

# TABLE OF CONTENTS

# Cyberattacks: new trends and a malicious ecosystem taking shape

*Occurring more frequently since the pandemic began, cyberattacks are now following new trends. Cyberattackers are also becoming more organised, dividing tasks between them to make attacks more dangerous.*

The cybercrime ecosystem seems to have undergone a transformation in recent months. Identified as an accelerator of the trend, the Covid-19 pandemic caused a net rise in cyberattacks. According to the French National Information Systems Security Agency (ANSSI), reports of ransomware attacks increased by 255% between 2019 and 2021. There is every reason to expect that this acceleration will continue.

### Ransomware attacks in the United States

- **30%:** the increase in SARs (Suspicious Activity Reports) submitted in the United States between 2020 and 2021 by American financial institutions on behalf of FinCEN, the American federal agency fighting online financial crime [FinCEN, 2021].
- **$590 million:** the total value of suspicious activity linked to ransomware in the first half of 2021 in the United States. For comparison, the figure was $416 million for the whole of 2020.
- **283%:** the rise in suspicious activity linked to ransomware between 2020 and 2021 in the United States.
- **$5.2 billion:** the total revenue generated by ransomware in the United States over 10 years (2011–2021).

Experts identify two qualitative changes: the first involves the nature of the attacks that take place, and particularly the new trends emerging. The second is specific to the organisation of the cyber ecosystem itself.

## Three new trends

*Big Game Hunting, RaaS, double extortion... 2021 was marked by the growth of three trends combined together.*

Emerging in 2018, Big Game Hunting uses attack methods and techniques that used to be the preserve of cyber-spying operations initiated by state-sponsored attackers. This type of approach, aiming to undermine a public or private organisation as a whole, has gained ground in recent months. Companies are targeted in advance and the attacks are carefully prepared. The data is often exfiltrated before being encrypted, and a ransom demand is sent to the top of the organisation. If they do not pay, the whole company is at risk [Securelist, 2020; Packetlabs, 2021]. The second trend is Ransomware as a Service (RaaS). Established by cybercriminals for the use of other cyberattackers, RaaS is based on the principle of SaaS, with a standard subscription including everything a cybercriminal needs to launch a ransomware attack. This model has the effect of multiplying both ransomware attacks and the methods used in these attempts to compromise systems. RaaS is characterised by its continuous mutations [CERT, 2021]. The third trend is double extortion. Arising in 2019, this is based not only on extorting a company's data but also on the threat of publishing it on a website or to the media [Threatpost, 2020; Twitter, 2021].

## A newly structured ecosystem

These practices are testament to a gradually emerging structure in the world of the cyberattackers themselves. The development of the techniques described above is now the preserve of organised groups, in which the division of labour is increasingly institutionalised. Sale of malicious code, supply of personal data, catalogues of compromised login details… The cybercrime ecosystem is increasingly developing its own products, with clearly identified buyers and sellers [Institut Montaigne, 2021]. We can now find services accessible online, such as infrastructure for hire for denial of service attacks or anonymisation. Cybercriminals are able to subcontract certain operations, including compromising their targets' IT systems using spam botnets. These distribution services infect their targets with phishing emails, spread malicious code within the IT systems and open up access to the compromised systems for their clients. There is a recruitment market for cyberattackers on the dark web, and targeted short-term missions with clear technical definitions can be offered. These structured and increasingly frequent attacks generate significant profits, in Europe but also in the United States, where suspicious activity associated with ransomware accumulated revenues of some $590 million between 2020 and 2021 (see box).

Faced with this sophistication, one of the challenges is technological, and involves mobilising artificial intelligence. As Olivier Gesny states in Revue Défense Nationale, AI is required to help reduce cyber risk, including improving cognitive capacity [Gesny, 2019].
This is precisely what concept code does.

### Health, education, IT outsourcing, local authorities: sectors in the firing line

While cyberattacks in general concern all geographical areas and can potentially strike any person or organisation, certain sectors are more heavily targeted than others:

- **Health**: hospitals and healthcare providers constitute a target for cyberattackers in the context of the Covid-19 pandemic [Bloomberg, 2020].
- **Education**: in the United States, this is the second most-attacked sector after local authorities [Enisa, 2020]. It is less subject to cyberattacks in France [CERT, 2021].
- **IT outsourcing companies**: this sector is particularly heavily targeted by cyberattackers.
- **Local authorities**: municipal councils, groupings of municipalities and district and regional authorities are targets for ransomware operators in France and worldwide. Town halls are particularly targeted for their low level of IT security, the presence of sensitive data and the problems caused by interruptions to activity.

## Cyberattacks in a few figures

Figures about cyberattacks – and especially financial figures – are difficult to estimate due to the multiplicity of sources, and the fact that ransom payments are often kept secret. Here is a listing of the main data available from specialist institutions:

- **2240** attacks were recorded per day in 2021. On average, an attack was launched every 39 seconds.
- Global cybersecurity spending exceeded **$1,000 billion in 2021.**
- The cost of the damage caused by cybercrime reached **$6,000 billion in 2021.**
- **9 people out of 10** in France have come into contact with a cybersecurity threat.
- In 2022, **6 billion people** are likely to suffer a cyberattack.

Cyberattack campaigns generate significant return on investment. According to a 2021 analysis by Institut Montaigne based on information from CERT-Wavestone, the net profit from a cyberattack conducted using RaaS (Ransomware as a Service) is between $500 thousand and $1.5 million. This translates into ROI (return on investment) of 232 to 880%.

*Sources : CERT, baromètre CESIN, Cyber'Occ, Cybermalveillance, Institut Montaigne, Kaspersky Lab, Ministère de l'Intérieur, Wavestone*

# Concept code: a revolutionary technology

*Inherent in viruses as in applications, computer code is at the heart of the new battle against malware. In response, thanks to AI (artificial intelligence), concept code is a breakthrough technology that focuses on the story told by the virus rather than the words and the grammar of which it is constituted. Explanations below.*

In 2022, the number of applications downloaded across the world is set to reach 258 billion. This momentum is growing steadily: in 2017, the number of applications downloaded was 178 billion, meaning the figure has risen by 45% in five years. As well as the underlying structural trend, this growth was strengthened by the context of the pandemic. In just a year, between 2020 and 2021, the market achieved growth initially forecast for two to three years.

## Growing difficulty in identifying malware

Naturally, this profusion leads to growth of a different kind, the rise of malware. Already significant, ransomware attacks on public and private organisations have increased significantly: between 2019 and 2020, the French National Information Systems Security Agency (ANSSI) recorded a rise of 255% [see previous pages]. As soon as a user runs a contaminated application, viral code is installed for each virus resident in the application. The virus remains on the device even after the user has closed the application, constituting an agent as active as it is discreet.

### AI, Machine Learning and Deep Learning

The AI engine developed by Frédéric Grelot and his teams of engineers is based on two elements: an algorithm able to read code and a continuously-enriched knowledge base. «We are training a Deep Learning algorithm every day to recognise any story told by a computer virus,» explains the engineer, a former executive at the French defence procurement agency and co-inventor of concept code with Cyrille Vignon. «In a way, the algorithm reproduces what a human brain does when it's learning to read – we train it to work on concepts, recognise them and compare them. During 2020, we carried out over 130,000 training sessions. In each training campaign, we feed billions of lines of computer code into our AI engine over four months of training, followed by two to three weeks of retraining. We make it read a story so that the AI can create its own knowledge base. Ultimately, all the concepts in malicious code will pass through its filter.» So far, the technology can identify 100 million concept codes, together with viruses that have been active for ten years.

Each application, each website and each video game has its own coding, i.e. its programming language. Viruses are no exception, except that the code they contain – the narrative, we could say – is malicious. The code remains essential for the virus to spread to other computers, tablets and smartphones.

It is increasingly difficult to identity the malicious payload hidden within an application, partly due to the growing use of artificial intelligence by the attackers themselves. Cybercriminals are making more and more use of AI to create their malware. Modified to varying degrees from one version to another, they create variants that present slightly different signatures. This enables them to escape the vigilance of conventional antivirus systems [see box].

### Understanding concept code through... children's fairy tales!

The analogy between concept code and the fairy tales we tell children helps explain the characteristics of this new technology. Because although fairy tales tell the same set of stories (from Goldilocks to Red Riding Hood), they are adapted for children in all cultures and many different languages. This is where concept code is special: it focuses on the story being told, rather than the words and the grammar used to tell it. This means the technology can identify scenarios that are 80, 90 or 95% identical in several types of malware, despite them being written differently. The story is at the heart of concept code, in the same way as the story is the basis of the emotions children feel in response to fairy tales and legends.

## Code conceptualisation: a stand-out technology

Faced with this increasingly sophisticated attacks, the lines of defence are tightening. In recent months, a new technology has made it possible to uncover this AI-based malware. Its name is concept code, or code conceptualisation. What is new about this technology, which operates using a combination of reverse engineering, machine learning and AI [see elsewhere], is its ability to identify malware that has never previously been detected. «We have developed a marker that stands out from other technologies,» explains Frédéric Grelot, one of the engineers *behind the invention. «Code represents functionality that we can uncover.»*

What is concept code, and how does this revolutionary technology work? To understand, we need to go back to the source, i.e. the code itself. Python, PHP, C, C++, Javascript, Java... Each of these programming languages is different

different in the same way as human languages are different, and each has its own grammar. «It's just like a foreign language,» explains Frédéric Grelot. "Through its code, a virus tells us a story. It's a bad story, but it's still a story. Malware uses the language of code, which is like the words and grammar of a book. What we need to understand is that the concept code technology focuses on the story the code is telling us, rather than the words it uses, because viruses often say the same thing using different words." This innovation is all the more important when we realise that current antivirus solutions focus more on the «words» than on the «story» they tell. Thanks to concept code, malware cannot leap over the safety barriers as easily as before. «Our marker is different from other technologies, and that's what makes concept code unique,» concludes Frédéric Grelot.

### Disassembling the code: the contribution of Reverse Engineering

Code conceptualisation is based on automatic Reverse Engineering. This involves disassembling the code, or translating it into understandable language, in order to identify its concepts. This process takes place in several stages, which are similar to the way a human learns to read, starting with learning the letters of an alphabet in order to form phrases and then learning the grammar. The end point of the learning process is the ability to assimilate the story being told, with all its different concepts and ideas.

# 10 million pieces of malware identified: one step on the road to Cyber Threat Intelligence

*Structuring the fight against malware and cyberattackers involves a significant effort to automate binary analysis. How is this done at present, and how does concept code play a central role in this work? A technical examination of the different stages of data collection and their concrete applications in identifying cyberattackers.*[1]

When a malware attack occurs, several central questions arise in terms of Cyber Threat Intelligence: who is the attacker? How are they acting? At this stage, it is vital to understand the intrinsic nature of the attack, but also, and above all, to pick up its weak signals. Time is precious to ensure the defence system can be deployed in order to limit or contain the damage. In this context, it is essential to capitalise on a set of data about malware and the families it belongs to in advance. Automation is now possible, and the concept code technology is helping to improve it.

## Data collection

Conducted on a large scale, the automation of binary collection makes it possible to analyse malware from end to end in record time. This approach involves several key stages, from collecting the malware to processing it via a pipeline and ultimately transforming it into concept code. The detailed malware inventory is produced principally from open sources of binaries. Linux repositories, open-source repositories such as GitHub, proprietary binaries, platforms such as Conan.io, Chocolatey, MalwareBazaar, VX-Underground and VirusShare... Reference sites are many and varied. Some list non-malicious software («goodware»), and others only malware, sometimes providing source code. The number of files accessible varies from 100,000 to over 10 million. In most cases, the viruses listed are less than three years old. However, some sites offer malware traceability stretching back ten years or more, varying in their ease of access.

The data collection pipeline consists of three stages: ingestion, tagging and finally transformation into concept code.

[1] *This page was produced using information from the following scientific paper: GRELOT Frédéric, LARINIER Sébastien and SALMON Marie, «Automatisation de l'analyse des binaires : de la collecte source ouvert à la Threat Intelligence», 16 November 2021, C&ESAR 2021 conference. URL: https://conf.researchr.org/details/cesar-2021/call-for-papers/14/Automatisation-de-l-analyse-de-binaires-de-la-collecte-source-ouverte-la-Threat-I*

## Data ingestion

The first phase, ingestion, is itself divided into several sub-steps: download, extraction, filtering and storage of the data. Download takes place via web scraping, using the Beautiful Soup platform. This involves finding download links in the pages of various data sources.

The extraction phase is specific to each data source. It allows us to extract additional information depending on the source. For example, files downloaded from the VX-Underground source platform allow us to identify each type of malware and the family of attackers it belongs to (APTxx). This information will be exploited during the tagging phase.

The next stage, filtration, has two goals. The first is to calculate the similarity of the identified malware to other malware in the database. This is determined using SSDEEP hashing. A second hash (SHA256) makes sure the download in progress is not from an archive or malware that has already been ingested, avoiding duplication.

Meanwhile, storage is essential. It must allow data of varying types to be stored and retrieved on demand, while also enabling the storage infrastructure to evolve in an agile

### Use case: proven links between the Babuk, Ryuk and Conti families

The malware analysis automation method described here has been applied to several use cases. The Babuk, Ryuk and Conti malware families have been passed through the filter of concept code and reverse engineering analysis. These viruses have struck a number of organisations recently. Emerging from the ransomware family, Babuk first appeared in December 2020, before its code evolved in May 2021. Seen for the first time in 2018, Ryuk has also seen its functions evolve: in October 2020, it was responsible for 75% of attacks in the American healthcare sector [CERT, 2021]. Conti, meanwhile, appeared in 2020 and has also struck many organisations, including the Irish health system in May 2021 (291 reported victims).

Applied to these viruses, the concept code correlation matrix has demonstrated links between several series of strains of the Babuk virus. In addition, the technology has shown that there was an identical development persona to Conti for certain samples of Ryuk. In particular, the compilation chronology allowed us to visualise an interweaving between Conti and Ryuk.

This information is extremely valuable for Cyber Threat Intelligence specialists: linking the threat to a specific sample supplements the malware database and guides the strategy of the incident response teams working within organisations.

way. We have chosen to store metadata in an Elasticsearch cluster and data on a Nexus server.

## Tagging the data and converting it into concept code

The tagging stage supplements the information specific to the malware and references it in Elasticsearch. Two types of taggers are used here: metadata and YARA rules. The first type uses the source metadata (tags from the virus collection platform, such as the malware family), while the second uses a series of rules to identify specific families. The last phase, the transition to concept code, takes place once all the binaries have been collected and identified. This is the transformation and conversion stage. It automatically correlates the malware and the families it belongs to via an asynchronous collection chain that makes it easy to resume processing, interrupting collection and restarting it later. Link searches for each source are carried out every 10 minutes, and tags are updated hourly.

This kind of automation is essential. It currently allows us to process 8.2 million items of malware, of which 3.3 million are tagged and 4.4 million are kept. This data collection is undoubtedly a step further towards Cyber Threat Intelligence.

### Concept code: a scientifically valid technology

The concept code technology was recently described in a scientific paper. Written jointly by Frédéric Grelot, Marie Salmon and Sébastien Larinier, the article compared two methods of binary analysis: manual analysis by an expert from the EPITA artificial intelligence engineering school, and automatic analysis by engineers from the start-up GLIMPS. Ultimately, the automated search and the concept code technology were validated by the manual academic approach. Conducted on a small scale (around fifty viruses), the comparison showed that automation using reverse engineering is not only reliable, but also much quicker than the conventional approach. This automation saves valuable time for defending organisations.

# Interview with Sebastien Larinier

## *Concept code: «I immediately realised the results were very interesting»*

## Concept code: «I immediately realised the results were very interesting»

*A lecturer and researcher at ESIEA and a member of the Digital Trust and Security laboratory, Sébastien Larinier recently discovered the concept code technology. Here he explains the benefits of this AI-based innovation and the objectives that can now be achieved as a result in terms of cyberdefence.*

### You work on virus analysis: what is the context of your research, and what does it involve?

**Sébastien Larinier:** I teach virus analysis at ESIEA, the engineering school for useful digital technology. It is one of 204 French engineering schools accredited to issue engineering degrees, and we are based in Paris, Ivry-sur-Seine and Laval. I work as a lecturer and researcher within the CNS laboratory (Digital Trust and Security). My work focuses on viruses and ransomware, and especially so-called nation-state attacks, also known as APTs or Advanced Persistent Threats. China and India are among the countries where I carry out research.

### How did you come to discover concept code?

**S.L:** Completely by chance. I was contacted on Twitter by Frédéric Grelot, one of the engineers behind the technology. He must have identified my core research, and he invited me to learn about the correlation matrix for Babuk, one of the viruses I work on and know best. Comparing the notes I had made with the results of the matrix, I immediately realised the results were very interesting.

### Why?

**S.L :** The case of the Babuk virus is actually a textbook case. I work on it rather like a geneticist might do with a human virus, looking for changes of sequence. This enables me to do what we call code monitoring, and see the different versions of the virus. In genetics, this would be known as phylogeny... The tool associated with concept code has the vital benefit that it can identify all the variant families automatically based on the statistics it generates. This saves time for researchers: we can now feed the virus families into a kind of grinder and automatically obtain all the variants, as long as they have not been modified by programs known as "packers". In addition, for the Babuk family, the tool highlighted the samples that provided the link between different versions.

### So the advantage of concept code is its automation?

**S.L :** Yes, but that's not all. The automatic dimension is certainly a major element of the technology. But so are speed, efficiency and security. Imagine: up to now, I did my calculations by hand. It was long and laborious, and the resulting fatigue can lead to errors. Concept code reproduces this work quickly for millions of viruses. What's more, concept code can highlight interesting, pivotal elements. Of course, the analyst still has to verify afterwards what all the virus families collected have in common, cluster by cluster. But at least once this check has been done, we can focus our work on a single «individual», a single piece of malware. We can then put forward hypotheses that will apply to all these individuals. As an analyst, that's great – it saves time and increases efficiency and security. The algorithm will never get it wrong. The algorithm just has to be well-designed. And I've seen that that's what happens with the concept code technology in the cases we've been able to work on.

### Ultimately, what possibilities does the technology offer in concrete terms?

**S.L:** Concept code has made artificial intelligence models efficient in analysing computer viruses and their similarities. It's a problem we have been faced with for years, but until now we haven't used AI, or only in an extremely complex way that rarely made it outside the laboratory. Concept code also tells us other things, within a virus family. You have to understand that cyberattackers use techniques of deception by changing the genetic part of the code. They use tricks that are specific to them. As a tool, concept code can demonstrate that identical strategies are used within different malware families. This means we can show that two families of virus may be the work of a single group of cyberattackers. We can identify the techniques they use and thus unmask them. Once we have understood that, we can organise an effective defence, such as generating an antivirus signature. We can block the malware earlier, and this can be done on a huge scale when whole families use the same process. This also applies to intrusion detection systems and EDR (Endpoint Detection and Response).

### Malware detection: a race

Speed is a major benefit when it comes to cybersecurity. When an attack occurs, it is essential to respond as quickly as possible, especially in terms of detection. «We currently see that algorithms can be effective at detection, but it takes time,» explains Valérian Comiti, R&D Engineer and Operations Director at the start-up GLIMPS. «If a defence solution has a doubt about a case, it will emulate the code to try to determine what is happening. It can take a variable amount of time, between a minute and an hour, to execute the binary. This is often long enough for the malware to spread through the system and cause significant damage.» In response, concept code has an advantage: it «reads» the code linearly and unpacks all the concepts it contains, with no exclusions. «Concept extraction is very quick, taking no more than three seconds,» continues Valérian Comiti. In other words, the concept code technology saves precious time by concentrating exclusively on the story told by the virus, rather than the words or the grammar it uses. «That’s the difference between emulating code and reading it. If we had to emulate and simulate execution in full, that would take time. By conceptualising the code from the first reading, we can summarise its contents and achieve detection speeds out of all proportion to what can currently be achieved.»

# Tomorrow: how can we defend ourselves against malware?

### The variant, a danger for all organisations

Variants constitute the real problem currently being seen in the field of cybersecurity. Out of the 2,240 attacks recorded every day (one attack every 39 seconds on average), the proportion of malware consisting of simple variants of native viruses is estimated at 99.9%. «The point of producing variants is that they can bypass traditional security barriers,» explain Frédéric Grelot and Cyrille Vignon, specialist cybersecurity engineers. According to some analysts, 230,000 malware variants are produced every day, generating massive quantitative growth.

The simple fact that the grammar of these viruses has been slightly modified is a guarantee of «success» against conventional lines of defence. An antivirus system is programmed to identify a «language» and a grammar, not to identify the story told by the code... «In fact, all the attacks that succeed today use variants: that's the very essence of their success,» point out Frédéric Grelot and Cyrille Vignon. Hence the need to flush out the malicious intent at the heart of the code's design – the malware's narrative.

Passing all variants through the filter of concept code technology is currently one of the only effective techniques. There is no doubt that malware production will develop further in the coming months, as cybercriminals adopt increasingly structured or even industrialised processes.

*What is the most appropriate system to defend against the cyberattacks that organisations are currently experiencing? We questioned two specialist engineers on the subject, examining the concept code technology and its evolution towards the notion of binary genealogy.*

Faced with the proliferation of attacks, multiplicity of vulnerabilities, diversity of technologies and unprecedented growth in malware, is there an optimum means of defence? According to Valérian Comiti, an R&D engineer specialising in computer security, the defence strategy to apply is just as complex and detailed as the attacks themselves.

## Multiplying security solutions

«*The best defence is one that involves multiple, integrated security solutions,*» explains the specialist. «*What is needed is a wide variety of defence technologies that all complement each other.*» This means there is a multiplicity of options: observing workstation activity using EDR, analysing inbound and outbound flows, concentrating on behaviour, a detailed analysis of what happens inside the flows... «*Some solutions excel at detecting strictly malicious behaviour, while others identify variants, like the concept code I'm working on at the moment. Again, there is no single* *solution: the solution is the proper coordination of varied lines of defence.»* But there is one certainty – technology is a key element of the defence that is currently being implemented within many private and public-sector organisations. *«Thanks to concept code, we have a stand-out tool that will create major problems for cybercriminals,»* emphasises Valérian Comiti. *«It's checkmate in the sense that the technology we use reads the code and immediately determines whether the system is infected with a malware variant.»* It is hard for attackers to get round organisations that use concept code: no binary can escape the gaze of a system armed with Artificial Intelligence. At the current stage of the battle being fought by many start-ups – including French companies – against cyberattacks, concept code offers an undoubted advantage.

## Binary genealogy, a new frontier

Will the same be true tomorrow? Frédéric Grelot is resolutely optimistic on this point. «*We are already seeing that two types of attacks are possible today,*» explains the École Polytechnique graduate in physics and computer engineering. «*The first case, which we see very often, is where a cyberattacker has modified an existing piece of malware. In this case, concept code will unmask the variant. This currently represents the vast majority of the attacks that occur.*» But what happens if the malware is a new type, with code that has never been seen before? «*This is where it is important for the defence system to be complementary, with EDR, for example, to identify malicious behaviour at a specific workstation. The new concept*

code that will then be generated will make it possible to examine the rest of the IT system with the fine-toothed comb of verification to see whether other machines are infected and neutralise the malware.» Links are gradually emerging between the different types of malicious software. Ancestors, descendants, lineage... The evolving work currently being done by R&D engineers specialising in cybersecurity has a direct connection to the concept of family relationships between malware, which could be described as binary genealogy.

We can draw a parallel with human or animal genealogy: the goal is to define relationships between families based on a genome that we can now decode. *«It's perfectly possible to reveal links between different concept codes to see whether there are family relationships between then,»* concludes Frédéric Grelot. *«That's what we plan to do in the future. Even if it's not an urgent task yet, we know it will be an avenue to explore if we want to take a step further in our research.»*

### How concept code adapts to the security environment of an organisation's IT systems

*«We work on a case-by-case basis, with no standard process, as every environment is different... We encounter problems, and we solve them.»* Jérémy Bouétard, a cybersecurity engineer and Chief Technical Officer at a start-up specialising in concept code, confirms that even with a variety of environments, it is not very difficult to integrate this new technology into an information system.

*«The solution is relatively simple to connect up given that our teams have automated and documented all the steps... The most difficult part is undoubtedly that the company's software has to be opened up. After that, it's just a question of running a script that will request the domain name and configure the server.»*

Once this operation has been completed, the organisation has a dashboard that can tell it in three seconds whether any pieces of malware are present. Some organisations want to go further, finding out which zones the malicious binary is resident in... *«In some cases, in big organisations, all these departments and divisions have to be coordinated. But I have to say we never get stuck from a technical viewpoint.»*

# GLOSSARY

### Concept code:

Code conceptualisation is a recent technology that involves automatically identifying the narrative of a piece of malware. Once it has been disassembled through reverse engineering, the code delivers up the story it contains, going beyond the numbers that constitute its words and their grammar. If the malware was a book containing a fairy tale, conceptualising the code would reveal the story. This process makes it possible to identify families of malware, given that 99% of malicious software consists of variants (variations) of an existing story. Concept code can thus be used to establish that a story is 5%, 10% or 15% identical to another, enabling the spread of malware to be stopped in its tracks. Code conceptualisation has already been tested in the academic field in recent years, but manually. The new element comes from the automation of this technology using Reverse Engineering, Artificial Intelligence and Machine Learning.

### Artificial Intelligence:

AI is a set of technologies that work together to enable machines to perceive, understand, act and learn with levels of intelligence tending towards that of humans. It includes several strands of technology, such as Natural Language Processing (NLP) and Machine Learning (ML). Artificial Intelligence can be either narrow (for concrete applications such as a virtual assistant) or general (i.e. «strong», capable of strategic, abstract, creative thinking in a variety of complex tasks). In the case of malware detection using the concept code technology, AI plays a central role by compiling all the families of malicious software identified over ten years or so. This represents millions of data points.

## Machine Learning:

Machine Learning is one of the themes of Artificial Intelligence, a subcategory of AI aiming to automate the process of creating analytical models. Machine Learning enables machines to adapt autonomously to new scenarios. One example of its use is the intelligent management of big data, which is a vital asset in malware research. Fundamentally, Machine Learning involves a computer examining data and identifying patterns in order to complete the task it is defined for (in this case, detecting malicious software).

## Reverse Engineering:

Reverse Engineering is a technique by which a system can be analysed in the opposite direction to its creation, starting from the end product and deducing its constituent methods and techniques. In the field of computer systems, this study and analysis can be applied to software, whether it is malicious (malware) or not (goodware). Software Reverse Engineering involves disassembling the machine code of malware to reconstitute the level of understanding provided by the original source code, using programming language instructions. The technique can be used to understand a program whose source code has been lost or is not known. This allows Reverse Engineering to identify the malicious content of a program.

## Correlation matrix (based on concept code):

A two-dimensional matrix gathering together the similarity values calculated for a set of binaries, comparing their concept code in pairs. The matrix provides a concise representation of their similarities and is used as a basis for analysing the links between pieces of malware from the same or different families.

# BIBLIOGRAPHIC REFERENCES

BLOOMBERG, « *Hackers Bearing Down on U.S. Hospitals Have More Attacks Planned* », 30 octobre 2020. URL : https://www.bloomberg.com/news/articles/2020-10-30/hackers-bearing-down-on-u-s-hospitals-have-more-attacks-planned

CERT, « *Etat de la menace rançongiciel à l'encontre des entreprises et des institutions* », sept. 2021. URL : https://www.cert.ssi.gouv.fr/uploads/CERTFR-2021-CTI-001.pdf

CERT, « *Le Groupe Cybercriminel TA505* », *10 février 2021 (version actualisée).* REDMINE : https://www.cert.ssi.gouv.fr/cti/CERTFR-2020-CTI-006/

CERT, « *Etat de la menace rançongiciels à l'encontre des entreprises et institutions* », 6 octobre 2021. URL : https://www.cert.ssi.gouv.fr/cti/CERTFR-2021-CTI-001/

DIGITAL SHADOWS, « *DarkSide : The New Ransomware Group behind Highly Targeted Attacks* ». 22 septembre 2020. URL : https://www.digitalshadows.com/blog-and-research/darkside-the-new-ransomware-group-behind-highly-targeted-attacks/

ENISA, « *ENISA Threat Landscape 2020 – Ransomware* », 23 octobre 2020. URL : https://www.enisa.europa.eu/publications/ransomware.

FINCEN, « *Financial Trend Analysis. Ransomware Trends in Bank Secrecy Act Data Between January 2021 and June 2021* », octobre 2021. URL : https://www.fincen.gov/sites/default/files/2021-10/Financial%20Trend%20Analysis_Ransomware%20508%20FINAL.pdf

FRANCE INFO, « *Cyberattaques : le nombre de piratages a quadruplé l'année dernière* », 16 février 2021. URL : https://www.francetvinfo.fr/internet/securite-sur-internet/cyberattaques/cyberattaques-le-nombre-de-piratages-a-quadruple-l-annee-derniere-selon-un-expert-en-cybersecurite_4299525.html

Gesny Olivier, « *Capter l'IA de demain au regard des enjeux de cyberdéfense* », Revue Défense Nationale, 2019/5 (N° 820), p. 38-42. DOI : 10.3917/rdna.820.0038. URL : https://www.cairn.info/revue-defense-nationale-2019-5-page-38.htm

GRELOT Frédéric, LARINIER Sébastien et SALMON Marie, « *Automatisation de l'analyse des binaires : de la collecte source ouvert à la Threat Intelligence* », 16 novembre 2021, salon C&AESAR 2021. URL : https://conf.researchr.org/details/cesar-2021/call-for-papers/14/Automatisation-de-l-analyse-de-binaires-de-la-collecte-source-ouverte-la-Threat-I

INSTITUT MONTAIGNE, « *Cybercrime : plongée dans l'écosystème* », 2021. URL : https://www.institutmontaigne.org/blog/cybercrime-plongee-dans-lecosysteme

Informatique News, « *Tableau de bord pour visualiser les cyberattaques à travers le monde* », 27 août 2021. URL : https://www.informatiquenews.fr/15-tableaux-de-bord-pour-visualiser-les-cyberattaques-a-travers-le-monde-72060

PACKETLABS, « *Qu'est-ce que la chasse au gros gibier ?* », 2012. URL : https://www.packetlabs.net/big-game-hunting/

REYNAUD Florian, « *Qui sont les hackers qui ont récemment paralysé une partie du système de santé de l'Irlande avec un rançongiciel ?* », Le Monde, 15 mai 2021. URL : https://www.lemonde.fr/pixels/article/2021/05/15/qui-sont-les-pirates-qui-ont-frappe-le-systeme-de-sante-irlandais-avec-un-rancongiciel_6080311_4408996.html

SECURE LIST, « *Targeted Ransomware : It's Not Just about Encrypting Your Data !* », 11 novembre 2020. URL : https://securelist.com/targeted-ransomware-encrypting-data/99255/

THREATPOST, « *Egregor Ransomware Threatens 'Mass-Media' Release of Corporate Data* », 2 octobre 2020. URL : https://threatpost.com/egregor-ransomware-mass-media-corporate-data/159816/.

TWITTER. @malwrhunerteam. 8 janvier 2021. URL : https://twitter.com/malwrhunterteam/status/1347458694053822464.

VADESECURE, « *Ransomware as a service (RaaS) : une activité illicite qui a désormais pignon sur rue* », 19 mars 2020. URL : https://www.vadesecure.com/fr/blog/ransomware-as-a-service-raas-une-activite-illicite-qui-a-desormais-pignon-sur-rue

GLIMPS

Web : https://www.glimps.fr/en/
Linkedin : https://www.linkedin.com/company/glimpsre/
Twitter : https://twitter.com/GlimpsRe
Email : contact@glimps.re