Cyber Data Sanitization: A Cyber Revival at the Heart of the Next Data Battle

Arnaud Le Dez

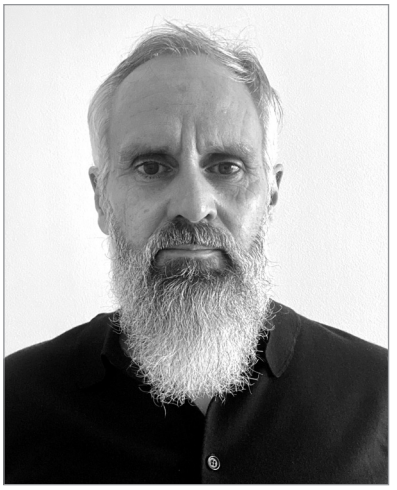
INTRODUCTION

We are only at the beginning of the history of cyber defense and the future holds immense promise. We have already seen significant advancements, and with the increased reliance on data, there is no end in sight. Foresight provides us with a unique opportunity to envision various potential cyber futures. It serves as a framework for crafting scenarios, which we can elucidate using methodologies, data analysis, and research. These futures become much more tangible when the key players and the factors that shape them are understood.

Among potential futures, the battle against data pollution emerges as a particularly promising prospect, thanks to the advent of new capabilities such as those described here and derived from cyber technology. This battle signifies not just an enhancement of cyber defense, but its evolution and expansion, paving the way for a new domain.

Data pollution is the degradation of the digital environment by data that can be considered as waste or a nuisance. These data can be naturally produced by digital systems for their operation or linked to human activities in the digital space. Data pollution is likely to affect the health of digital systems and the quality of processing, leading to degradation or interference with operations in cyberspace. We are in a familiar universe here, cyber-attacks can be a form of pollution.

U.S. Army Cyber Command refers to data pollution as a "data rationalization" problem.[1] If the cybercommunity does not solve the problem of data pollution, we will not be able to pursue our activities with the same efficiency and by developing new artificial intelligence capabilities. Just as cyber encompasses all digital activities, this fight against data pollution involves all systems, including command and control systems and other areas of warfare and intelligence systems where data optimization is a constant objective.



Arnaud Le Dez works for Capgemini in Paris as a consultant on cyber strategy and governance for C-level executives and CISOs of major companies. He was a French Army officer for 28 years, with responsibilities in electronic warfare, cyber defense, and intelligence. He has over six years of experience working with the U.S. Army during several missions in the Middle East. During the last ten years of his military career, he held various responsibilities in cyber defense operations in the French COMCYBER. He is an associate researcher at the conflict transformation division of Saint-Cyr Coëtquidan Military Academy Research Center and is the author of a book on the tactician level in cyber defense "Tactique Cyber, le Combat Numérique," published in January 2019 in *Economica*.

Data analysis is at the core of cyber defense and cybersecurity tactics. Once processed, data yields a cognitive effect that benefits both defenders and attackers. A major concern within cyber is the over-processing of data. A novel tactic within cybersecurity aims to sanitize the digital battlefield for improved detection of nefarious acts and the enablement of counteractions.

The battle against data pollution strives to optimize data processing. The issue is as much about optimizing processing as it is about reducing digital costs and limiting the fog of data war to enable operating in a more visible data environment. It curbs certain excesses, especially uncontrolled data generation from various sources like cybersecurity, intelligence, artificial intelligence, faulty systems, saturation attacks, and more. It morphs cyber defense into an ecological capability that aligns with today's societal and battlefield challenges. This innovative and beneficial application area has the potential to rejuvenate the cyber defense operational environment, broadening its scope with optimized data use at its core.

The battle against data pollution is requires a partial shift in the application of cyber tools, which will detect pollution as an assault on our systems. Sanitation has always been a cyber concern. Now, it takes on an even greater role: the need to sanitize data will accompany the defender through all stages of digital conflict.

In addition to unlocking a new market via a fresh class of cyber-based tools, the battle against data pollution will also enhance cyber defense and the efficiency of command-and-control systems across the battlefield. It will allow cyber defenders to concentrate on genuine attacks by cleaning up the digital environment. This cleanup will also free processing capacity on tactical mission command systems that are currently inefficient in data processing. The battle against data pollution aids in equipping our tactical formations with more potent cyber tools.

We all need to know when we create and/or use systems that generate polluted data. We must recognize it and remedy it for the sake of operational efficiency. Data pollution is becoming our fog of war. The intent of this article is to show that we can combat data pollution by reusing knowledge and tools we use in cyber in what can be called the digital domain environment. This concept applies to beyond just the cybersecurity practitioner to all commanders who rely on high-quality data to make decisions.

THE CYBER MODEL: ITS LIMITATIONS AND EVOLUTION

Cyber defense derives from numerous developments in the evolution of cybersecurity. Originally, the reliability and availability of systems were initially dictated by advancements in electronics and protocols, which form the backbone of our current digital capabilities. The 1990s saw an expansion of interconnections via the Internet and the elevation of security of information systems to a higher priority. This was particularly due to the emergence of new tools and structured governance. The Y2K bug presented an opportunity to rectify past mistakes and establish contemporary security policies.

As tools evolved and it became apparent that mere protection of our systems was inadequate, we had to operationalize the cyber profession in an environment riddled with ongoing attacks. This operational approach, centered on the need for real-time actions to defend against attacks that are truly combat, facilitated the transition to the present-day cyber defense. The time from detection to action has been significantly reduced, thanks to new capabilities for generating and processing cyber data, enabling a swift response that increasingly anticipates the unexpected and mirrors the interactions found in military combat.

The current cyber defense model is based on the generation of an astronomical amount of data in systems dedicated exclusively to cyber use, followed by real-time analysis to detect anomalies that could signify attacks. During this operational phase, cyber intelligence has become indispensable for parameterizing tools to detect and characterize adversaries, regardless of their location. It is necessary to know and understand an adversary's tactics and weapons when developing pre-validated automatic responses.

The cyber community is using its tools to generate increasing amounts of data[2] from a growing number of sources, driven by the fear of overlooking the slightest indication of nefarious actions in a world where our adversaries are constantly adapting. This data accumulation presents its own set of challenges. Not only is it time-consuming to store and process on a massive scale, but it is also difficult to extend our tools into highly constrained environments. The extension of cyber into new environments, towards all digital objects of everyday life (or weapons) and into increasingly contested digital universes, is one of today's most significant challenges.

The industrial environment that supports much of the world's critical infrastructure is currently at the epicenter of commercial digital transformation and is simultaneously the target of

increasingly disruptive and costly attacks. The defense sector views it as an important part of the field of conflict: it is a source of cyber intelligence, but it also needs national-defense level counter-measures. This necessitates a rethinking of cyber defense tools.

We are increasingly using smaller and more autonomous digital devices that often have limited data processing and storage capacity. Anything our military or government employees use, both in the course of their jobs and in their private lives, could be targeted by our adversaries. These systems include digital industrial or military equipment, cars, satellites, weapons with digital systems, watches, or personal fitness devices.[3] Such systems either need to be designed for defense or assessed for the risk they could pose if targeted by our adversaries. However, the deployment of our cyber tools is often not planned and often not possible for these devices. They are often closed systems, limited in their processing capacity, constrained in their interconnections, and not designed to accommodate outside cyber tools.

ARTIFICIAL INTELLIGENCE AND DIGITAL TWINS: THE PIONEERS OF TODAY'S TRANSFORMATION

Artificial intelligence and digital twins[4] are at the forefront of today's transformation, further amplifying the presence and importance of data. For instance, we can use digital twins, virtual representations of physical or information technology systems that is updated in real-time, to apply intelligent simulation algorithms and predict anomalies indicative of future attacks or improve those we are planning using AI. In the defense sector, the combination of a digital twin with artificial intelligence will pave the way for predictive combat in the cyber domain, for both the offense and defense, enabling military planners to better synchronize cyber operations with operations in the other warfighting domains. Using AI and digital twins in a cyber fight will lead to a cleaner digital environment and reduce complexity, allowing commanders a clearer view of battlefield, both in the present and the future. Clean data will make operational planning easier because there are fewer unknowns. Cleaner data and a clean cyberspace environment contribute to the predictability of cyber actions, reduces the fog of cyber warfare, and makes cyber operations planning more relevant.

Therefore, we must question the limitations of our models across all battlefield systems. By their very nature, they restrict cyber action in certain digital facilities; artificial intelligence and digital twins could potentially exacerbate this problem. Increasing processing capacities to meet this challenge is costly, as are the potential battles lost, should our digital combat systems fail. The volume of data generated across our current systems is colossal. The amount of data pollution or, at best, single-use data, is significant and poses challenges for both energy conservation and resource optimization. These challenges are amplified in constrained or contested environments.

Cyber defense will need to evolve and mandatorily involve a more selective approach to data: less data, or higher quality, with a better understanding of its power and limitations. We thus

need to create different digital blueprints to apply various cyber data processing schemes, thereby increasing the probability of detecting attacks.

This is a global phenomenon that does not just affect cybersecurity. Reducing data pollution will become a major objective: this problem is ubiquitous. Even though cybersecurity and cyber defense are currently noise generators within the digital landscape, they are also part of the solution, as we shall see below.

THE CYBER BATTLE AGAINST DATA POLLUTION

There is a need to repurpose cybersecurity tools to clean up data beyond just that which is deemed malicious. Often cybersecurity tools analyze data and generate a response that can be simplified to "good data or bad data." The analysis grid is based solely on harmfulness, in terms of whether the data are part of a cyber-attack. Cyber tools have the capacity to analyze and process data in all systems, with a logic that is both autonomous and integrates centralized control and increasingly automated action. The detection, command, and response chain, often called the XDR-SIEM-SOAR chain, is completed by antivirus, firewalls, intrusion detection and protection systems (IDS/IPS), and other tools.

Today, processing criteria within the cyber community are only concerned with security. We can broaden this limited focus and transform our cyber tools and cybersecurity using an ecological approach that includes the fight against data pollution. This will enable us to create a new range of tools in the marketplace, and to go beyond the current limits of cybersecurity and its extensions.

The use of cyber tools to combat data pollution can first be envisioned by modifying the criterion of harmfulness. For example, firewalls filter data and can be extended to block all forms of pollution. Detection cyber tools (EDR/NDR/XDR) and antivirus systems analyze data, with the ability to detect and clean them up (or quarantine them). Security Operations Centers (SOCs) steer the entire process and manage resource optimization thanks to a form of intelligence that seeks to characterize what pollution is and what the tactics, techniques, and procedures (TTPs) of these polluters are. All cyber systems can potentially be used to combat data pollution, often after a few simple modifications, possibly involving some intellectual and technical adjustments.

The most challenging aspect is the characterization of data pollution, a task that requires the continuous updating of pollution indicators. We are going to have to invent Data Pollution Threat Intelligence (DPTI), just as we have Cyber Threat Intelligence (CTI). We need to build an evolving dictionary for characterizing pollution, just as we have built an evolving dictionary for characterizing cyber-attacks. This characterization is as complex as the definition for what constitutes malicious data, and this lies at the heart of the cyber problem.

We must also remember that data that are polluted today may not have been in the past or will not be in the future. Pollution may be obvious and constant, but it can also be temporary, making it essential to have an ongoing ability to clean up the data. AI could help us by

generating identical data with the same processing purpose, while at the same time eliminating pollution. Cleansing is an essential component in the fight against data pollution, and here again, the tools used could be derived in part from other cybersecurity efforts.

It is important to bear in mind that cyber tools will not be the only solution. With the increase in sensors across the world as well as the increase is the use of deception, data pollution has the potential to be a much more important phenomenon than cyber-attacks. There is a need to scale up, and to promote low-cost digital hygiene solutions. The solution will include new standards and controls along with a reactive alert network. In this new environment, there is a place for innovative software companies to detect, alert, and cleanse data.

CYBER OBJECTIVES IN COMBATING DATA POLLUTION

The battle against data pollution offers numerous benefits. The most evident is the decluttering of our systems by limiting our data to only what is relevant. Navigating directly to pertinent data and examining it from various perspectives will be a crucial challenge in ensuring that cyber does not get submerged in digital info-obesity.

The advantages for cyber are clear. A system with less data, both in flow and in storage, is naturally easier to monitor with cybersecurity tools. It is always simpler to detect an attack on a clean system than on a polluted one.

Considering that some cyber-attacks also constitute a form of pollution, these new tools will aid in curbing the spread of the threat. For instance, spam and network scans can be categorized as pollution. While network scans are necessary on operational networks, it should be ultra-compressed and encrypted and when complete the results should then be placed in a virtual “garbage can” or, in case cyber intelligence or cyber forensics analysts need them, in a separate database, away from operational networks. Reducing these data would enable cyber analysts and tools to focus their energy and effort on the more complex attacks. It would also allow cybersecurity algorithms to operate on technically more constrained systems, thereby extending the defended areas.

Cybersecurity is poised to branch out into new areas, including the industrial world and the numerous digital devices we use daily. These devices were not designed to accommodate the astronomical amount of data that our cyber models generate today. This effort to limit data and processing will therefore be a prerequisite for deploying our cybersecurity on certain low-processing-power equipment.

Pollution control can be seen as a cyber hygiene measure, using a combination of tools and techniques that need to be kept simple and inexpensive. Cleaning up our data is not just a cyber issue, but a more global or enterprise-level case of data management and data optimization - including the data we use in cyber security and cyber defense. The development of a new field is a business opportunity, with new financing and a new way of approaching a market that is much larger than cyber, and which will broadly enhance our security and defense.

AI AND THE ONSLAUGHT OF CYBER DATA

Artificial intelligence (AI) will undoubtedly lead to significant advancements in cyber analysis capabilities. This is the route currently favored by hardware and software manufacturers.[5]

AI is heavily dependent on the volume of data it requires for learning. It also generates a substantial amount of it.[6] A recently emerging phenomenon is the synthetic data production to feed other AI algorithms, whose output, often called ghosts, is then fed into the datasets used to develop the next AI algorithm in sequence until the desired outcomes are achieved. These ghosts are a dream where data appear plausible but do not actually contain real data in the context of the answer; some would call this fake data. This result is the so-called hallucination effect where AI developers and their resulting algorithms lose touch with the reality they are supposed to interpret. Used this way, AI is producing data pollution on a scale that is proportional to the growing power of AI. This hallucination is the equivalent of shooting oneself in the foot, which will overwhelm systems and produce incongruities triggering massive malfunctions.[7]

To maintain a healthy digital environment, some Artificial Intelligence will need to be equipped with tools that can detect pollution and clean up the data they produce.[8] This battle against pollution generated by AIs would be a method of control for AI production environments to maintain system availability. Ideally, the battle against data pollution would address the integrity of results by proactively seeking to eliminate the ghosts contributing to future malfunctions. We cannot work on dream data that will pollute our appreciation of reality and the quality of treatments. This will be even more essential when dealing with mission-critical systems, such as weapons systems, cars, or satellites. This mirrors the reasoning in cybersecurity, where we strive to define the normal behavior of a system as a method to detect abnormalities that could potentially be attacks.

NEW CYBER FRONTIERS

This new vision of cybersecurity should generate new interest for deployment in digital universes that have been neglected until now, or in highly confined environments. The global significance of combating data pollution is similar to that of comprehensive security, while simultaneously reaching audiences who are sometimes resistant to the concept of security. This implies that this new category of tools must liberate itself from this supervisory bond. Viewing cyber from an ecological perspective can attract new followers to a reputedly depleted field, possibly even gaining traction at the level of decision makers, hence contributing to the cyber defense cause.

The fight against data pollution helps to deploy cyber systems in highly constrained situations. Some digital infrastructures remain limited in the processing of data, as evidenced by operationally deployed military forces. Efforts to reduce the volume of unnecessary data are crucial given the military's increased use of interconnected digital equipment. The mastery of data contributes to the agility of system usage, whether in the military or in business.

In the military spirit of economy of means, less data equates to fewer flows, less storage media, and data processing with reduced energy consumption - and a reduced aging of the equipment. This should also contribute to the Low Data movement, which aims to accomplish the same tasks with less data. Undoubtedly, in a high-intensity conflict scenario involving massive use of cyber weapons, the ability to act in a degraded digital battlefield with tools that consume less data will help us to preserve freedom of action.

CONCLUSION: IT IS A QUESTION OF DEFINING EVIL DATA

Feedback from cyber experience is invaluable. We can morph cyber to generate an ecological capability that fits into a much larger global dimension. In cyber, it is difficult to define what, exactly, is an attack, characterizing it, and then blocking the malicious data. Everything depends on the characterization of cyber evil.

We face the same difficulty in the fight against data pollution, except there is also a challenge regarding the persistence of this characterization. Data that are not polluted today may become so tomorrow and, more likely, they will become obsolete. We are witnessing a constant evolution of the threat. Good, timely, and legitimate data today, maybe be deemed evil or obsolete tomorrow, but then later be relevant as a mission focus shifts. We are already familiar with this phenomenon. The difficulty is relative because the fight against data pollution basically uses the same thinking, the same organization, and the same tools as we have in cyber. The transformation is mainly based on the definition of pollution, on the ability to define the TTPs for handling data pollution.

This characterization of pollution is much more than a technical or organizational challenge, it is also of a philosophical/political nature. The data used in combatting cyber threats has similar philosophical dilemmas. If undesirable data were labeled as polluted data, any software developed under this principle could become a formidable censorship tool. Some countries have adopted this philosophy while others declined.

The battle against data pollution poses the exact same issue. At its heart lies the question of what we mean by polluted data, and whether this includes the informational level. The challenge here arises from the abuses that can occur at this level. We already have reflexes such as parental filters, but we must bear in mind that the subject is well known, and there are as many different answers as possible, often linked culturally. However, we need to be cautious, because these tools can also be used to suppress information that someone for selfish reasons would not want to see propagated.

While particularly useful as a military tool in the context of information warfare, we must remain aware that countering data pollution requires a degree of vigilance and control by policy, the virtue of democracy, and the guiding finger of the market.

NOTES

1. George I. Corbari, Neil Khatod, John F. Popiak, and Pete Sinclair, "Mission Thread Analysis: Establishing a Common Framework in a Multi-discipline Domain to Enhance Defensive Cyberspace Operations," (April 2024). *The Cyber Defense Review*, Volume 9, No. 1, pp 37-54. https://cyberdefensereview.army.mil/Portals/ 6/Documents/2024_Spring/CDR_CDRV9N1-WEB-Spring-2024.pdf, 46.
2. Mohammed M. Alani, "Big data in cybersecurity: a survey of applications and future trends," (January 6, 2021). *Journal of Reliable Intelligent Environments*. https://doi.org/10.1007/s40860-020-00120-3. 7, 85–114.
3. Dayton Ward, "Fit to be Spied: Fitness Trackers and OPSEC Risks," (March 9, 2018). *The NCO Journal*, Army University Press. https://www.armyupress.army.mil/Portals/7/nco-journal/docs/Fitbit.pdf.
4. Korolov, Alex, "The cybersecurity challenges and opportunities of digital twins," (December 6, 2022). *CSO Online*, https://www.csoonline.com/article/574179/the-cybersecurity-challenges-and-opportunities-of-digital-twins.html.
5. Mitchelson, Deryck, "Learn about the Double-Edged Sword of AI in Cybersecurity," (October 27, 2023). *World Economic Forum*, https://www.weforum.org/agenda/2023/10/the-double-edged-sword-of-artificial-intelligence-in-cybersecurity/.
6. Orf, Darren, "A New Study Says AI Is Eating Its Own Tail," (October 29, 2023). *Popular Mechanics*, https://www.popularmechanics.com/technology/a44675279/ai-content-model-collapse/.
7. Alcaraz, Anthony, "When AI Distorts Its Own Reality: The Dual Threats of Model Collapse and Source Bias," (November 11, 2023). In *Medium*, https://ai.plainenglish.io/when-ai-distorts-its-own-reality-the-dual-threats-of-model-collapse-and-source-bias-13e4da42fd10.
8. Quentin Bertrand, Avishek Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gide, "On the Stability of Iterative Retraining of Generative Models on their own Data," (September 30, 2023). Conference Paper, *The International Conference on Learning Representations*. Vienna. https://doi.org/10.48550/arXiv.2310.00429.

DISCLAIMER

The views expressed in this work are those of the author and do not reflect the official policy or position of Saint-Cyr Coëtquidan Military Academy Research Center, the French COMCYBER, or the government of France.