DARKTRACE INDUSTRIAL

# Triton 2.0 & The Future of OT Cyber-Attacks

### Key Takeaways

- Leveraging machine learning and AI tools is now a feasible option for malware creators
- Traditional security tools will fail to detect a next generation, 'Triton 2.0' attack, that can blend in to the native environment
- Whilst offensive AI tools have limited access to data, defensive AI technology has full visibility of the digital environment, giving it the upper hand in a new cyber arms race
- Artificial intelligence is essential to defend OT systems

## Introduction

Over the last few years, cyber-attacks on Operational Technology have increased rapidly in frequency and scale. As geopolitical tensions are reflected in cyberspace and attacker technologies become more advanced, the cyber-threat to critical infrastructure and other key operational systems is now front and center of national security concerns. There is a new frontline in cyber defense where protecting against increasingly sophisticated attacks and anticipating future developments in attacker tradecraft is crucial.

This white paper evaluates the development of known OT attack campaigns, and the wider progression of malware, in order to identify trends and extrapolate future scenarios. Special focus is given to the emergence of AI and machine learning techniques, which have revolutionized cyber defense, and will become even more critical as we look into a near future where machine learning is also used by attackers. The grave dangers of weaponized AI are particularly acute in the OT space, and are likely to lead to a form of cyber arms race where only the best AI system will win out.

> **The grave dangers of weaponized AI are likely to lead to a form of cyber arms race.**

# High-Profile OT Attacks

Six major cyber campaigns against Operational Technology have been made public, from the infamous Stuxnet attack in 2010, which first demonstrated that operational control system networks were viable targets, through to Triton in 2017, malware which took down critical safety systems in the industrial control units and halted the operations of at least one facility.

| Campaign | Year | Malware Effort / Cost | Threat Actor Effort / Cost | Common Mechanisms | Distinctive Mechanisms |
|---|---|---|---|---|---|
| Stuxnet | 2010 | High | Low | Command & Control | USB stick & File infection<br>Four zero-day exploits<br>Prepared control- system specific attack |
| Havex | 2014 | High | Low | Spear phishing<br>Command & Control | Watering-Hole<br>OPC Enumeration |
| Steel Mill | 2014 | Low | High | Spear phishing<br>Commodity IT components<br>Command & Control | Sabotage crippled control system |
| BlackEnergy | 2015 | High | High | Spear phishing<br>Commodity IT components<br>Command & Control | Reconnaissance leading to system specific sabotage attacks |
| Industroyer | 2016 | Low | Low | Spear phishing<br>Commodity IT components<br>Command & Control | Custom network scanner<br>Specific PLC attack<br>OT protocol scanning |
| Triton | 2017 | Low | High | Spear phishing<br>Commodity IT components<br>Command & Control | OT protocol scanning<br>Specific control system reprogram<br>Safety systems compromised |

> **Since the Stuxnet virus first demonstrated the vulnerability of Cyber-Physical Systems, we have witnessed numerous destructive attacks that have caused serious and costly damage.**

# Key Attack Trends

## Commodity malware

The most striking progression is the shift from bespoke malware designed to hit specific OT targets (Stuxnet, Havex), to the use of commodity malware aimed at generic IT systems, which append a small OT-specific module (Industroyer, Triton).

This change has drastically reduced the cost and range of expertise required to create effective, OT-targeting malware, and facilitated the progression of OT threats from a niche concern to a central part of the cyber-threat landscape.

## Command and Control (C2)

C2 channels are a consistent feature of advanced malware, and were used in all six attacks. C2 is used for updates, exfiltration of data, adding new modules and capabilities, and sometimes allowing humans manual control.

Developments in this area include hijacking established websites to act as the destination server, as well as the increased use of encrypted but common internet protocols such as HTTPS, as this often avoids inspection at the network border.

More recent C2 can wait until legitimate internet traffic is occurring in order to more effectively blend in. Darktrace has observed malware launched in a wide scattershot manner being later taken over by a human threat actor and controlled manually, having apparently landed in a network of specific interest.

Malware innovation that partially or fully removes the need for C2 would pose a significant challenge for current detection methods. Only Stuxnet had any significant lateral movement capability without its C2 connection, showing how malware could operate autonomously if it was designed in advance with extremely detailed knowledge of the target OT network.

## Spear phishing & watering-hole attacks

While spear phishing is the most common initial vector to get malware past firewalls, other methods include watering-hole attacks where a trusted third-party website is compromised, and malware delivered physically through infected media, such as a USB drive.

Employee security training is paramount to reduce risk, but users cannot always successfully avoid compromises, while some will act intentionally. Furthermore, on acquiring new software updates, organizations are effectively forced to outsource part of their security perimeter to the software vendor.

## IT threats in OT environments

Organizations have recently realized significant cost savings by using generic IT hardware with specialized software in OT environments, rather than developing unique OT hardware. However, over time this has meant that nearly all the common features – and thus vulnerabilities – of IT networks have been introduced into OT control systems.

Attackers have responded to this convergence by using generic IT malware components for OT-specific attacks, leading to increasing cases of OT networks being affected by malware never intended or specialized for them, such as the WannaCry ransomware. This attack showcased the extent to which OT systems relied on their connected IT networks, a topic addressed in the scope of the EU NIS Directive legislation for securing critical networks.

The sensitivity of OT to generic malware lowers the bar to entry for less technically sophisticated attackers. This raises the prospect of a new attack model that holds physical operations to ransom, in a similar way that ransomware holds data files to ransom.

Traditional enterprise networks are currently a more obvious target for organized criminals seeking quick monetization, as the development costs of the attack are significantly lower compared to attacks targeting OT. But this will shift quickly if generic IT attacks can be easily ported into OT environments. Recent trends in IT malware are therefore very relevant to the future of OT attacks.

# Key Hallmarks of a Sophisticated Cyber-Attack

### Persistence

Sophisticated attackers aim for persistence: a long-term foothold in the target environment, in spite of possible efforts to remove the malware. This may be achieved by using secondary infections on a single device or by rapidly spreading to multiple devices – exact methods keep changing. Log manipulation is a recent technique that targets traditional SIEM log-analysis detection approaches. By blocking or altering logs before they leave the infected device, the relevant events are never seen to be processed and the attack goes unnoticed.

The key to mitigation is the visibility that a security team has. If the team can see the entire infection, it can be remediated. If they can only see part of the infection, persistence mechanisms are likely to succeed.

### Polymorphism

Attack techniques for avoiding up-to-date anti-virus and IDS signatures have significantly improved in recent years. There are free websites that test malware files against all current anti-virus programs, so creators can be certain their efforts will be undetectable by signatures on release.

Software components that incorporate a polymorphic malware factor are easily purchasable, meaning that the malware changes every time it spreads. This makes developing a signature that can reliably detect the morphing malware extremely difficult. File-less malware, which hijacks already installed software for its own purposes, also bypasses most of the visibility of anti-virus programs.


### Monetization

While nation states do not rely on monetizing their malware, criminals have been held back by the difficulty of converting cyber compromises into untraceable currency. The anonymity of cryptocurrencies has made this process vastly easier, and the wide use of both ransomware and cryptomining malware have followed.

Cryptomining is of particular note because it has created a direct link between stealth and profit for criminals. With nation states already keen to remain undetected for as long as possible, hiding is one of the main targets of development. Malware is being designed to blend in with existing activity as much as possible.

# Future Attack Developments: AI-Powered Autonomous Malware

The deployment of machine learning and artificial intelligence in other fields of software means that leveraging these technologies is now practical for malware creators, and there are several clear ways it can be used to strengthen their efforts.

To blend in with the environment, an attacker must gain the best possible understanding of that environment. Today's most advanced malware can facilitate this process. Darktrace has already discovered attacks that leverage basic machine learning techniques to understand how an infected device normally communicates and therefore when and how it should be active to appear as similar as possible.

In the future, a highly effective use of machine learning will be to train malware in optimal decision-making. Cyber defense is suffering from a skills crisis, and this is no less true for the threat actors, who are reliant on experienced hackers. Supervised machine learning can transfer the skills of the best malware operators directly into the malware itself.

Greater autonomous ability within the malware will allow it to delay establishing a C2 connection. Having to maintain C2 is a point of weakness that can reveal the presence of an attack. Trained malware can operate independently until, for example, it is able to communicate with an OT control system. Establishing C2, performing OT reconnaissance and exfiltrating the results can then be completed extremely rapidly, far too fast for humans to mitigate the threat even if it was spotted immediately.

Both unsupervised learning to blend in and supervised learning for decision-making also have applications in the OT-specific payload part of industrial-targeted malware. Compared to devices operated by chaotic humans, the behavior of individual devices in an automated environment is usually more repeatable and could be learned and profiled effectively through machine learning techniques.

If we revisit the real-world OT attacks analyzed above, we can extrapolate potential evolutions of the malware and tooling used in these campaigns. Machine learning will be most relevant for campaigns designed to sabotage the control system in a way that leads to catastrophic failure. These attacks will be especially effective if both the control system and the safety system can be compromised in a way that avoids detection and a premature process shut-down. Assuming this approach, we can speculate about how AI techniques would be used in a future revision of similar malware.

> **Supervised machine learning can transfer the skills of the best malware operators directly into the malware itself.**

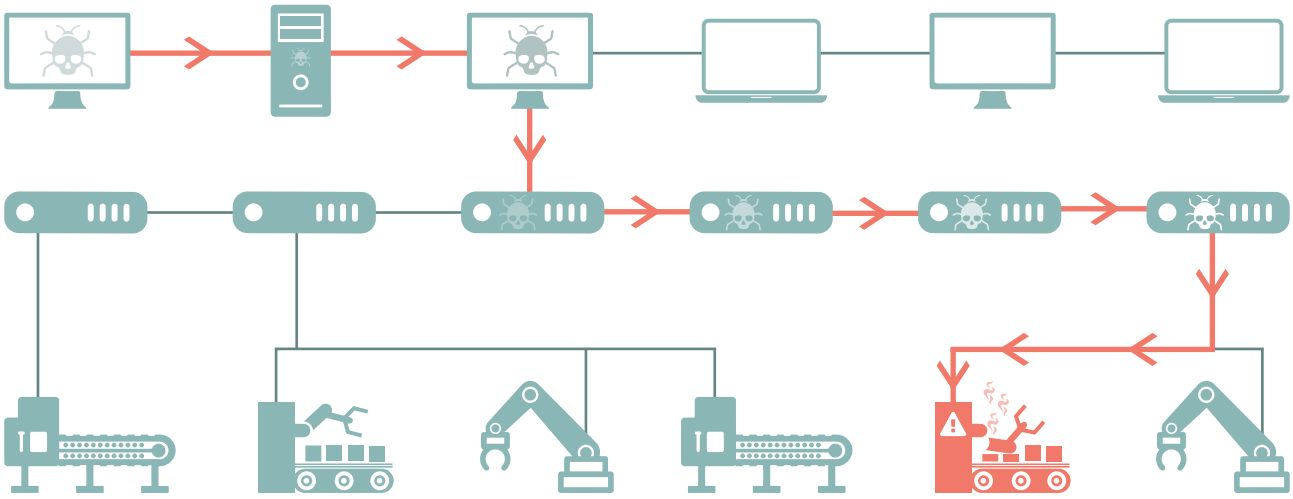

**Fig 1:** AI-enabled malware is able to autonomously find the optimal path to its ICS target.

# Defending Against 'Triton 2.0'

**An otherwise extremely sophisticated and effective campaign, the 2017 Triton malware attack was undone by accidentally stumbling into a non-obvious 'trip-wire'. A redundant pair of controllers in the safety system failed a validation check of the application code, triggering a shutdown.**

**Checks and detection mechanisms like this are inherently unknowable to attackers, and avoiding them is the primary challenge in systematically compromising an industrial control system.**

**AI will significantly improve the chances of a Triton-like attack succeeding at every stage: reconnaissance, access and persistence. Let's examine what a Triton 2.0 would look like.**

## Stage One: Reconnaissance

Malware designed to operate in safety systems can benefit from AI through mitigation of the need for command and control. The more local decision-making the malware is capable of, the less user input required. The Triton framework required operators of the malware to manually trigger its functions through scripts. Although in this case the attackers could operate at this stage without detection, we can imagine an AI-equipped version operating without command and control, perhaps only calling back at the end of the reconnaissance phase.

Achieving command and control is especially challenging from within highly secure zones, and detecting C2 channels is currently one of the best ways of identifying and preventing these attacks. Reducing reliance on C2 would radically increase the difficulty of detection with traditional tools.

During reconnaissance there are potential actions which the malware would perform that might trigger detections

- Simple IP based scanning to identify target hosts
- Mapping network topology to identify gateways and firewalls
- Protocol-specific identity requests
- Download of PLC programming for further analysis.

Any of these steps can trigger detection by either violating unknown detection rules in place on the environment, or by deviating from expected behavior for the compromised device if anomaly detection is in place. An AI-enabled framework could include functionality to analyze network communication from the compromised host and attempt to passively construct a network map based on the (limited) observed network traffic.

If multiple implants were achieved in the environment, these compromised systems could form a distributed peer-to-peer observation network in an attempt to piece together a more comprehensive picture of network traffic than would be available from one host alone.

The end goal of this passive reconnaissance would be to train machine learning algorithms in order to determine the 'likelihood of detection' and a 'likelihood of benefit' score based on the current level of understanding of the environment. These scores could then be utilized in decision-making logic to weigh risk of detection against the value of possible next steps in active reconnaissance to minimize accidentally tripping alerting systems.

By observing examples of legitimate communication between supervisory devices and controllers, the malware could output generic request packets that match observed formats and send these to additional targets, blending in and avoiding detection from simple content inspection and rule-based systems.

### Stage Two: Access

The biggest risk factor when accessing a controller is running into unknown dependencies and relationships, as with the Triton campaign. With AI, intelligent classification of targets based on behavior and possible dependencies between them could be used to minimize this risk. Armed with knowledge from passive and active reconnaissance, the trained classifiers could determine a risk score associated with the attacking of a particular target given the current state of the control system and the location of the implant.

For example, reprogramming of a PLC from a particular host may violate logical zoning requirements. Attempting to directly access a safety device from a non-authorized workstation may be possible at the network level due to insufficient segmentation, but may still trigger a rule or anomaly based alert.

Understanding common communication paths could allow the mapping of conduits and plausible behavior types throughout the network. A distributed framework that leverages multiple compromised hosts could select a host to launch the attack by calculating associated risk scores and acting to minimize them. If no course of action was found that satisfied minimum risk requirements, the malware could re-enter reconnaissance mode to attempt to find alternate paths and spread laterally to compromise additional pivot points.

### Stage Three: Persistence

Both Triton and previous OT attacks have regularly polled infected controllers to check on continued persistence. This behavior is unsubtle and could be avoided by the use of AI modules. Future malware could assess the likelihood that persistence has been lost by attempting to identify possible reprogramming and update events in the environment. Based on this information, intelligent decisions could be made to only check the status of controllers based on environmental changes that make it likely that persistence has been lost.

**Many of these steps would have been performed manually by the operators of Triton. An AI-enabled Triton would both be more effective at avoiding detection through this process of intelligence risk modeling, and be able to operate without the need for continual C2.**

# The Machine Fights Back

Malware is becoming much smarter and more adept at hiding and evading defenses. But machine learning used defensively is able to counter such attacks and starts each encounter with major advantages.

The level of assessment needed both to blend in or to detect requires a high level of visibility and access to data. Offensive AI attacks must use multiple implants and communicate peer-to-peer to build a sufficient picture of the activity across the control system. In general, a defensive system will have full knowledge and access to data across the whole environment, as opposed to a limited set of footholds. This difference in vantage point makes the training of detection algorithms in situ far easier than the training of the competing malware. With the wrong tools, the complexity of watching everything all of the time becomes a sea of data that hides an attacker. With the right tools, it instead becomes the advantage that the defender needs.

Darktrace's unsupervised machine learning evolves from having no knowledge at all, based purely on the network traffic and data that it sees. It does not bring in assumptions from elsewhere about how the network or any given device within it should function. As the monitoring is out-of-band, it cannot be manipulated or erased as logs can. This wide and complete visibility also strongly counters persistence mechanisms that like to live in blind spots.

As every network is unique, every Darktrace deployment is unique. This means that it is not possible to test malware against Darktrace in advance to see how it responds. Through probabilistic mathematics, Darktrace is able to become gradually suspicious that something unusual is occurring – it does not have to make instant black or white decisions about a given event – so malware trying to be 'low and slow' is not avoiding detection any more than rapid and blatant activity.

In contrast, rule-based and simple whitelist defenses are highly susceptible to basic learning malware techniques. Some of the most effective current malware 'lives off the land', misusing the legitimate tools that it encounters rather than bringing in its own. A very simple policy of only repeating observed legitimate actions will completely avoid simple detections. Only machine learning that considers a great deal of additional context can recognize unusual use of permitted actions.

External communications such as C2 and cryptomining transactions are often relatively easy to identify compared to the normal activity of a device, whether that is compared to its own history or to its peers. Ultimately, the attacker is doing something that the normal business does not, and this difference cannot be completely hidden. Comparisons with other devices allow Darktrace to identify even threats that pre-date its installation, despite having never seen a specific device behave any differently.

When malicious activity is fast (for example when ransomware reaches a networked file share), even instant detection is not sufficient because human response times are too slow. For these threats, only defenses that can react at machine speed will be able to stop them from accomplishing their goal, and to this end, Darktrace Antigena, the world's first Autonomous Response technology, is able to execute a targeted and proportionate action, containing threats in real time before they cause material harm.

> **Only machine learning that considers additional context can recognize unusual use of permitted actions.**


**Fig 2:** Darktrace's unsupervised machine learning delivers full visibility of the industrial environment

# Conclusion

Attacks against Operational Technology have taken advantage of general trends in malware development and attack commoditization, and no longer require significant custom development. This means that OT networks are firmly on the map for attackers, and made attractive due to the outdated IT systems that they often rely on. Indeed, custom OT-specific payloads only need to be attached to the very end of the kill chain and are usually the simplest parts of the attack code.

A major shift to the use of artificial intelligence in such attacks is set to fundamentally disrupt OT systems further, with malware now able to contextualize its surroundings and adapt accordingly, without relying on manual C2 connections. As a result, attack campaigns are becoming fast, efficient, and highly-targeted, whilst the risk of detection is reduced at every stage of the attack lifecycle.

> **Signature-based malware detection is dead. Cyber security needs a quantum leap forward. It needs to rely on machine learning-based artificial intelligence.**
>
> Senior Fellow
> Institute for Critical Infrastructure Technology

Whilst these AI attacks are antiquating legacy security tools relying on signature methods, they can be effectively detected by the use of Cyber AI on the defense side, thanks to the technology's ability to understand unique environments, and mathematically model what is 'normal' and what is 'abnormal' activity at any point in time. Indeed, Cyber AI defense already catches machine learning-enabled, polymorphic attack code on industrial networks in various industries, and alerts security teams who can in turn respond.

But in this new reality of cyber-threat, human teams' response time is quickly becoming inadequate. The near future of OT cyber security must rely on AI to not only detect the AI-powered, autonomous malware that we must be prepared for, but also to take measured actions, in some cases, to curb the activity immediately. AI must now fight AI to not only find otherwise undiscoverable threats, but also respond to those threats in real time.

## About Darktrace

Darktrace Industrial is the world's leading AI company for cyber defense. Created by mathematicians from the University of Cambridge, Darktrace's Industrial Immune System technology uses AI algorithms that mimic the human immune system to defend industrial networks of all types and sizes. In an era where OT and IT are increasingly converging, Darktrace's technology is uniquely positioned to provide full coverage of both enterprise and industrial environments. By applying advanced machine learning and AI, Darktrace Industrial defends Critical Infrastructure across the world, and is relied upon by leading energy providers, utility companies and manufactures to secure their ICS and SCADA environments.

## Contact Us

North America: +1 (415) 229 9100
Europe: +44 (0) 1223 394 100
Asia-Pacific: +65 6804 5010
Latin America: +55 11 97242 2011
info@darktrace.com
darktrace.com/industrial
@darktracel