Prevent: Security through Adversity*

A Darktrace Discourse Paper

February 8, 2022

'If you know the enemy and know yourself, you need not fear the result of a hundred battles. If you know yourself but not the enemy, for every victory gained you will also suffer a defeat. If you know neither the enemy nor yourself, you will succumb in every battle.'

'Attack is the secret of defence; defence is the planning of an attack.'

(Sun Tzu, "The Art of War" [8])

Abstract

The quotations above succinctly (and somewhat poetically) describe and emphasise the inextricable link between offensive and defensive activity. Although applicable to all areas of adversarial interaction, this short paper focuses on the ever-increasing requirement for augmentation of commonplace defensive measures with broad-spectrum adversarial simulation and/or emulation within the cyber security landscape. While all components of the Continuous Feedback Loop – prevent, detect, respond and heal are of great importance in establishing a robust defensive posture, particular focus will be granted within this paper to the first phase - *prevention*.

1 Introduction

As destructive cyberattacks against business and government organizations escalated in recent years, interest has grown in the fields of vulnerability assessments, penetration testing and red teaming aimed at improving the cyber-centric defences of organizations. Contemporary surveys, however, indicate that almost 30% of organizations do no form of adversarial assessment. The remaining 70% noted that even a relatively mature and well-resourced blue team will more often than not, fall-short of defending their critical assets from the red team [1]. While there are obvious benefits for organizations to conduct adversarial assessments, many organizations have yet to venture into this space for multiple reasons:

- **Resources** - A respectable red team consists of a highly talented and qualified group of individuals with a relatively rare set of skills. The demand for such services is high while the aforementioned individuals are in short supply. This results in a single red team exercise cost ranging between $40,000 and $80,000. [6]

*The author (who has chosen to remain anonymous) is based at the Darktrace AI Research Centre, Cambridge, UK and has 7 years of experience in the automation of complex cyber-centric processes with specialization in the offensive domain. The author also holds a PhD in Astrophysics from the University of Cambridge.

[1]Data collated from a Darktrace Roundtable of its CxO council members, comprised of CIOs and CISOs in all major markets (EMEA, APAC and the Americas)

- **Psychology** - Organizations yet to experience the misfortune of a cyberattack may find it difficult to justify the costs of a red-team exercise. Furthermore, with the red team claiming victory more often than not, the blue team and its leadership may experience a sense of inadequacy when they are unable to successfully defend the organisation. Finally, and not surprisingly, the internal security team may feel uneasy when an individual or group (particularly a third-party) have gained information that may enable the compromise of their network.

The issues outlined above must be addressed if organizations are to adopt adversarial practises and ultimately improve the security of their assets.

In general, cost reduction of complex processes carried out by highly skilled individuals can be achieved by pooling and automating the expertise to allow for deployment at scale. However, automation of such processes is by no means straightforward and there will likely always be a requirement for humans to drive the research into new attack vectors. Potential approaches to automation of these processes are explored in brief within this paper and in greater detail within future publications.

The psychological barriers to implementation of adversarial practises, require a mass-shift in community thinking, which, according to recent surveys appears to already be in motion and likely stems from the logically undeniable defensive advantages that come as a result.

The proceeding section aims to outline some of the axiomatic principles associated with preventative security, with the primary focus of taking on the mindset of an attacker in order to better prepare one's defences.

2 Elementary Principles

To effectively prepare for a task, one must train as close to (or exceeding the difficulty of) the real task as possible without adversely affecting the trainee. In military spheres, this is generally referred to as "Mission Specific Training". Unfortunately, whether due to lack of resources or psychological aversion, this *modus operandi* is generally inadequately practised within the field of Information Security, despite being extremely effective.

The proceeding subsections aim to provide the user with a set of four, elementary, guiding principles to assist in creating or developing a robust defensive posture.

2.1 Know Your Enemy

As alluded to in the opening quotations of this paper, understanding your adversary is vital to mounting a strong defence. Intuition would suggest that one must first contemplate the desires of their adversary - money, information, destruction, chaos? However, within the Information Security sphere, the assets associated with achieving these goals are generally speaking, quite similar. For example, a customer database containing large quantities of personally identifiable information is an enticing target for a malicious actor interested in money, information, destruction or chaos. Instead, greater focus should be applied to understanding how one's adversary might attempt to realise those desires. What tactics, techniques and procedures (TTPs) do they commonly utilize?

At a fundamental level, an attacker aims to identify vulnerabilities and exploit them to achieve their goals. Modern day organizations can be most generally categorised as a fusion of both **man** and **machine** working in synergy to accomplish certain outcomes.

Skilled adversaries will recognise the existence of vulnerabilities that the general populace are often blind to. When considering the methods by which an adversary may exploit our vulnerabili-

ties, it is critical that we cast away the legal and moral boundaries which the majority of us abide by within our daily lives, for they are unlikely to be adhered to by our adversary.

Fig. 1 outlines the variety of vulnerabilities which an adversary may attempt to exploit during an attack. It should be noted that the underlying concepts of software and hardware vulnerabilities can be readily extended to humankind. All people have vulnerabilities, some more universal than others and to varying degrees resulting from an evolutionarily developed code-base. This "code", developed via natural selection to ensure survivability and the passing on of genetics, left weaknesses. These double-edged swords which, more often than not, will save, but also have the potential to be utilized to manipulate, weaken or destroy.

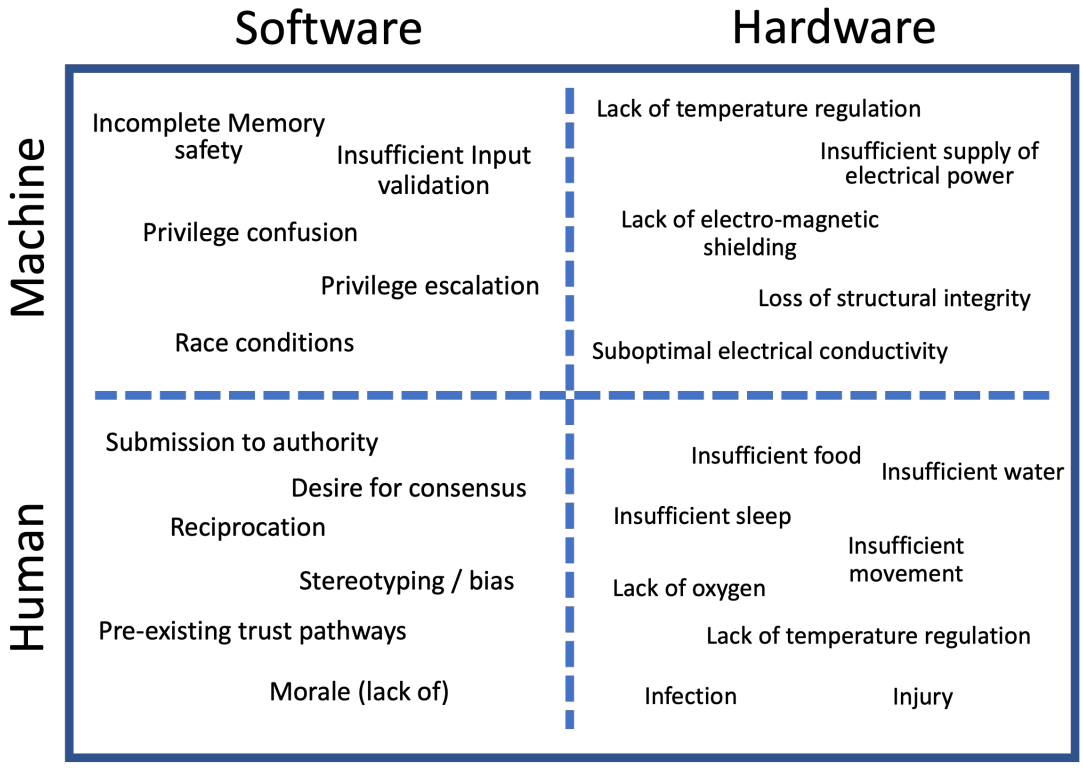

Figure 1: The vulnerability matrix of a typical, modern organization.

A short list of somewhat unconventional (and unsavoury) exploitation of both human and machine vulnerabilities which assisted in compromise of the target network is given below. These are sourced from personal experience or internationally reported events.

- During an exercise, a red team operative cloned the target organization CEO's voice using readily available, open-source machine learning software. A variety of phrases were synthesised to deal with potential questions from internal IT services. The attacker successfully feigned a broken company phone scenario in order to justify request of a VPN bypass code.
- In mid 2020, a Russian national known as Egor Igorevich Kriuchkov approached a Tesla employee via a mutual acquaintance, offering the employee $1 million in exchange for the deployment of malware into the Tesla network. The target was groomed over a period of weeks prior to the final request being made. In a testament to Tesla's strong workplace culture, the employee reported the approach to his employer and the FBI, enabling the arrest of the suspect during his attempt to escape back to Russia. [7]

- Stuxnet is a malicious computer worm first uncovered in 2010 and thought to have been in development since at least 2005. Stuxnet targets supervisory control and data acquisition (SCADA) systems and is believed to be responsible for causing substantial damage to the nuclear program of Iran. Stuxnet functions by targeting machines using the Microsoft Windows operating system and networks, then seeking out Siemens Step7 software. Stuxnet reportedly compromised Iranian PLCs, collecting information on industrial systems and causing the fast-spinning centrifuges to tear themselves apart. [2]

- A malicious actor identified the head of blue team for target organization via social media. Personal details were successfully correlated with publicly available breach data - matches were found for a user account associated with AdultFriendFinder. Contact details of close family members and company peers were subsequently identified. Ransom email was sent on the evening prior to initiation of attack in order to degrade efficacy of the defence via lack of sleep and elevated stress levels.

- On the morning of April 16, 2013, a team of gunmen, using rifles, opened fire on the Metcalf Transmission Substation, severely damaging 17 transformers. The first bank of transformers, riddled with bullet holes and having leaked 52,000 US gallons (200,000 l; 43,000 imp gal) of oil, overheated, whereupon PG&E's control centre about 90 miles (140 km) north received an equipment-failure alarm. [3]

These examples will hopefully broaden the reader's understanding of the creative (and sometimes immoral) strategies which their adversaries develop and execute in order to achieve their goals.

For a more conventional but very well maintained and established list of techniques used by red teams and malicious actors - the MITRE ATT&CK framework is an excellent resource. [4]

2.2 Know Yourself

The remaining component of the opening quotation refers to knowing oneself. Applying this concept to a typical modern organization emphasises the requirement of visibility for both human and machine infrastructure, for you cannot defend what is yours if you do not know it exists. At first glance, this may sound straightforward, but modern-day organizations are highly complex and few have 100% accurate and up-to-date knowledge of their digital assets. Both of these aspects must be addressed in order to optimise one's ability to defend and requires a multi-faceted approach for comprehensive coverage.

Firstly, an organization should aim to exhaustively identify all assets for which the blue team are responsible. Working from the outside in, "Attack Surface Management" (ASM) tools will typically take a domain or keyword as a start point and iteratively identify all associated domains, subdomains, IP blocks, TLS certificates, open ports and services.

Care should be taken to not exclude the human components of the organization, especially as socially engineered phishing emails still remain the dominant entry points for malicious actors [5]. These ASM tools provide a well-structured and (hopefully) comprehensive view of externally facing infrastructure which should (in principle) bear a strong resemblance to that observed by a malicious actor once their initial reconnaissance phase is complete.

Network monitoring tools are (relatively) simple to deploy and provide real-time insight into the internal aspects of one's organization with behavioural characterisation of network traffic providing invaluable information regarding discovery of new devices and the probable roles those devices are carrying out. Furthermore, network monitoring provides excellent "big-picture" visibility during an

attack due to the simple fact that it is incredibly unlikely for an attacker to enter the target network at their desired final destination and will thus rely on the network to move laterally towards their objective.

Although generally requiring greater effort to deploy across a large number of devices, endpoint monitoring solutions can provide highly detailed information regarding the active processes which a network monitoring tool may be blind to.

Lastly, but by no means least, monitoring of cloud assets with the plethora of "Software as a Service" (SaaS) applications which are now so widespread within modern organizations' day-to-day practises is absolutely critical.

Once an asset register has been completed to the best of an organization's ability, it is then vital to ensure and centralise real-time logging of data from those assets. Without this, high-velocity threats such as ransomware may have successfully completed their task before you have even been made aware any abnormalities or activity linked to the compromise. Naturally, real-time data flooding in from all known assets with a detail level sufficient to identify threats will require processing power far beyond that of a human to extract and analyse information required to identify a cyber attack.

As previously stated, one should not exclude the human components of an organization from this principle. Visibility and "logging" provided by your employees is critical to the determination of their vulnerability state. This "logging" could be in the form of user-driven reporting of suspicious activity encountered or information regarding personal issues which may affect overall susceptibility to exploitation by a malicious actor.

In summary, this section aims to emphasise the requirement of thorough knowledge regarding which assets to defend and obtaining up-to-date information from these assets for further, timely analysis. Without this, one cannot even begin to prioritise the allocation of resources 2.3 or deploy countermeasures 2.4 appropriately.

2.3 Prioritise

No organization exists with infinite resources. Therefore, one should logically aim to prioritise the allocation of said resources in order to maximise efficacy. In order to prioritise intelligently, a "risk matrix" approach which aims to assign impact and probability to situations of interest can be utilized. One would then naturally assign high priority to high impact combined with high probability scenarios, low priority to low impact combined with low probability scenarios and medium priority to the remaining cross terms.

Within the realms of simulating/emulating a cyber attack, the calculation of probability/likelihood and impact level associated with an attack scenario can be extremely complex. Furthermore, accurate estimation of these quantities relies heavily on the principles outlined in the preceding sections 2.1 and 2.2 to be carried out to a high standard lest we fall foul of uncertainty propagation.

A valuable tool to assist in prioritisation of resource allocation is Attack Path Modelling (APM). APM utilizes techniques from Graph Theory, representing organizational networks as directional, weighted graphs in order to identify the shortest (lowest resistance) lateral movement paths to key assets. This should ideally be fuelled with a fusion of data sources, including (but not limited to) network, endpoint, Active Directory, SaaS and email data.

It is important to take a moment to distinguish between Simulation and Emulation as the two concepts have already been referenced in this paper.

- A **simulation** is typically detached from the real world; the output of a simulation is not directly connected to the thing it simulates. For example, an aircraft simulator does not

actually fly, and the pilot is not actually communicating with a real air traffic controller. A simulation usually has the goal of testing or predicting some real-life process in a safe environment; because the simulation is disconnected from the real world, the risk of negative consequences is significantly diminished. This naturally extends into the cyber security domain of simulated attacks using a vulnerability model which is based upon characteristics of the observed environment with zero potential for adverse consequences to the target organization.

- An **emulation** by contrast, has the goal of taking the place of the real thing: for example, a conventional red team exercise in which humans are requested to carry out a cyber attack (within a given scope) is considered an emulation. While closer to reality, it has the added drawbacks of potentially causing real disruption and/or damage to an organization.

If accurately modelled, APM simulations can provide the blue team with invaluable information, enabling the progressive neutralisation of potential attack routes without the risk of disruption to normal business activity which can sometimes result from emulation (as opposed to simulation) of lateral movement. However, circumstances may arise under which information may be insufficient in order to reliably infer the potential for lateral movement. Under such circumstances, where simulation confidence is low, emulation should be attempted with care in order to ensure accurate modelling and ultimately prioritisation for defence of one's assets.

Once more, attention should be paid to the human component of an organization. Prioritising the defence of certain individuals can be achieved by assessing the level of external exposure and criticality to the ongoing activity of the organization. Additionally, user-awareness training can be prioritised according to the measured susceptibility of individuals to emulated social engineering attacks.

2.4 Deploy Countermeasures

Consider an adversarial system in which equal resource is provided to each actor. During each interaction, actors will utilize their respective resources in order to neutralise the opposition. If at each interaction, measures can be deployed which result in asymmetric resource deduction in favour of the defence, the resources of the attacker will tend to zero faster than those of the defence, thus preventing the attacker from achieving their goals. While an attacker may have perceived advantage due to elevated readiness state, lack of morals or requirement to abide by the law it should be noted that they too, do not have infinite resources.

It is often stated that security and practicality are inversely correlated. However, under some circumstances, one may identify scenarios in which countermeasures can be deployed such that asymmetric resource deduction between adversaries can be achieved. For example, maintaining up-to-date patch management significantly increases the resource requirement of the attacker to continue along the same attack path as it would require the discovery of a zero-day vulnerability. However, one should take heed and not entertain the fallacy that fully patched software is invulnerable. Relatively simple analysis of Common Vulnerabilities and Exposure (CVE) data [1] can be used to derive metrics such as high CVSS vulnerabilities per unit time for different pieces of software. Such metrics allow the blue team to either replace, or prioritise the defence of devices running specific software which (even if fully patched) hold an intrinsically elevated vulnerability status.

It is important to remember that attacks rarely begin inside the target network. An adversary will typically carry out an extensive phase of reconnaissance against the target organization using both passive and active techniques. For example, during this reconnaissance phase, an attacker

might register a fake LinkedIn account and pose as a new starter of the target organization (this was made especially easy inside the predominantly home-based onboarding period during COVID-19 lockdowns). LinkedIn currently does not employ any method of verification regarding your status as an employee of the organization you claim to work for. One would then typically proceed to converse with current employees, feigning ignorance while steering towards the acquisition of useful or exploitable information such as contact telephone numbers of internal IT services, payroll etc. alongside organizational idiosyncrasies which may be utilized in future as crude methods of authentication. In an effort to counter this activity, a simple script was produced, which monitored for the presence of new employees associated with the target organization on a regular basis and compared this list with an up-to-date employee inventory. Mismatches were then automatically reported via LinkedIn's fake profile reporting mechanism.

The deployment of such a countermeasure required a relatively small amount of effort on behalf of the blue team but effectively neutralised an entire mechanism for the active gathering of target data by the attacker. The creation of new fake profiles costs the attacker time and effort, which if repelled, results in the asymmetric deduction of resource in favour of the defence alluded to at the beginning of this section. One should aim to erect countermeasures at the earliest stages of the attack if possible - particularly if they are likely to cost the attacker more dearly than the defender. Deploying these countermeasures at the earliest stages of an attack has the added psychological benefit of preventing the attacker gaining "momentum". Although dependent upon the attacker's mentality and underlying motives, generally speaking, if one's attempts to exploit vulnerabilities are repeatedly thwarted at an early stage (it is often assumed that difficulty will increase as one progresses), this may serve to deter the attacker entirely.

Finally, as touched-on towards the end of Section 2.2, processing the volumes of data associated with comprehensive, real-time logging from an organization's assets is far beyond the analytical power of any human being. In the near future, advancement of modern technology will result in greater integration between machine and organization, yielding even larger and likely more complex data feeds. Unfortunately, adversaries will, in parallel receive similarly advancing technology and will refine their techniques to become more automated, stealthy and fast-moving. In light of this, data feeds will likely require greater detail in order to identify potential threats, nearer real-time delivery and processing in order to respond in a timely fashion. Current data processing requirements combined with the predictable advancement of technology make machine learning driven analytics the only (currently known) viable, future-proof solution to identifying non-standard threats. This does not mean that signature-based detection mechanisms are without their uses. Such systems are extremely useful for detecting threats with low recall but high precision at low financial and computational cost. However, even a relatively unskilled adversary should be able to avoid detection by these signature-based systems. In contrast, machine-learning (both unsupervised and supervised) driven security solutions force even the most skilled attackers to expend considerable time and effort attempting to baseline the behaviour of the compromised device before attempting any activity.

3 Conclusion

The principles outlined in Section 2 provide the reader with a progressive approach to optimising their defensive posture with a particular emphasis on the utility of adopting the mindset of one's adversary and ensuring asymmetric resource deduction at their expense. Although touched upon briefly, the specifics of analytical processes associated with technical solutions will be examined in greater detail within future publications alongside a series elaborating upon the Continuous

Feedback Loop covering the functions of Prevent, Detect, Respond and Heal loop.

In summary, if a defender understands the methods of the attacker 2.1, has an awareness of their own assets 2.2, appropriately prioritises the allocation of their resources 2.3 and deploys countermeasures 2.4 such that the attacker will expend resource at an asymmetric disadvantage with each attempted advance, the blue team will be victorious.

References

[1] Cve details. https://www.cvedetails.com/.

[2] Stuxnet. https://en.wikipedia.org/wiki/Stuxnet, 2010.

[3] Metcalf sniper attack. https://en.wikipedia.org/wiki/Metcalf_sniper_attack, 2013.

[4] Doug P. Miller Blake E. Strom, Andy Applebaum. Mitre att&ck: Design and philosophy. https://attack.mitre.org/docs/ATTACK_Design_and_Philosophy_March_2020.pdf, July 2018.

[5] Security Intelligence. State of the phish: Ibm x-force reveals current phishing attack trends. https://securityintelligence.com/posts/state-of-the-phish-ibm-x-force-reveals-current-phishing-attack-trends/.

[6] Mitnick Security. What should you budget for a penetration test? the true cost. https://www.mitnicksecurity.com/blog/what-should-you-budget-for-a-penetration-test-the-true-cost, January 2021.

[7] SecureWorld News Team. Inside an fbi sting: The ransomware gang trying to bribe your employees. https://www.secureworld.io/industry-news/fbi-sting-the-ransomware-gang-trying-to-bribe-employees, August 2020.

[8] S. Tzu. *The Art of War*. Dover Military History, Weapons, Armor. Dover Publications, 5th Century BC.