

Adversarial Machine Learning: The Coming Legal Storm

Peter R. Stephenson ¹
School of Law , University of Leicester

Abstract. The intersection of technology with the law always has been slow and often fraught with confusion. Lawyers do not speak the language of technology and, conversely, technologists are uniformly ignorant of the law's twists and turns. The purpose of this paper is to bring the legal community in closer tune with the technical community without forcing the legal community to undertake an in-depth education in machine learning (ML).

A key area of confusion is jurisdiction in cyberspace. There are, essentially, two camps. One side says that cyberspace is a domain of its own without laws, governance or enforcement. The other side claims that, in most cases international law as it stands is sufficient for managing legal aspects of the Internet.

The second is correct to a point. However, there are emerging areas of technology that appear to defy that position. One of the most dangerous and difficult – both theoretically and practically – is the area of autonomous malware. This paper discusses that challenge and proposes some approaches to meeting it.

We begin with some technical definitions and explanations to frame the legal issues properly in the technical world. We then address jurisdiction in cyberspace generally followed by demonstrating a standardized way of looking at the Internet through the lens of the law. Next we present a hypothetical attack of the type envisioned. Finally, we pose some possible solutions to the problem of autonomous malware and software-driven – as opposed to human-driven - attacks.

¹Corresponding author: pstephen@cdfs.us

Please note that this paper is focused upon United States law and the state of technology in the US. However, the author has cited an African journal article that suggests that the issues stated here are present throughout the world (see footnote 35).

Keywords. cyberspace, cyber law, jurisdiction, international law, private law, cyber definitions, autonomous malware, Internet

Introduction – the Lawyer as Data Scientist

The core of autonomous malware is artificial intelligence (AI). The core of artificial intelligence is machine learning (ML). These two concepts imply that something – malware in this case – can operate without human guidance. Granted, human guidance is present in the form of the software programs that operate the object under control, but once that software takes over, the object operates autonomously managed only by the software (software-driven).

Take, for example, the autonomous vehicle. The automobile is given sensors, software and control systems that allow it to know where it is, what is around it and what the driver's wishes are. It then goes to work to accomplish the driver's mission taking its environment into consideration and responding accordingly. The software is in control.

In this paper we are concerned with machine learning and, more specifically, adversarial machine learning (AML²). Before we tackle the legal issues spawned by ML and AML in cyberspace, however, we must simplify and explain the language of the intelligent machine.

The first question we must ask is, "What is ML"? Next, we need to understand what we want the machine to learn and how it learns it. Finally, we want to understand how the machine – in this case an autonomous malware, or "bot" – uses those data. Once we answer those foundational questions we can move on to the legal issues, secure in the knowledge that we understand – at least at a fundamental level – the technical ones.

There are several definitions of machine learning but they all boil down to the same thing:

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn

²Lawyers in some jurisdictions typically use the acronym for Anti-Money Laundering but we use it for Adversarial Machine Learning

and improve from experience without being explicitly programmed³.

We want our machine to learn how to identify certain things such as its environment, input data, output data and how to manage those data. In the case of an autonomous bot, for example, we might want it to learn what kind of system it is entering (banking, manufacturing, medical, etc.), what data it is supposed to locate (payment cards, medical records, trade secrets, etc.) and what it is supposed to do with those data (steal, delete, modify, etc.).

To do these tasks the bot needs to learn what the things it is seeking look like. Now things get complicated because every banking system, for example, is not set up the same way nor does it, digitally, look the same. So we have three ways our bot can learn what its target looks like.

The first way is supervised machine learning. In supervised machine learning you provide the bot with a list of data that look like the target and tell it to go find these. Consider the old way of learning the multiplication tables: constant repetition and rote memorization. That is supervised training⁴.

Unsupervised training, on the other hand, is achieved through observation. We observe and draw conclusions from what we observe. The more we observe, the more our conclusions are refined⁵.

There is a third learning mode, semi-supervised learning where you start with some data as in supervised learning and deduce the rest through observation, basing your observations upon what you already know from the supervised portion. The least dangerous of these three is, arguably, unsupervised but it can be quite a bit slower than supervised since it must learn from scratch.

Another issue with machine learning is error. You will see that in a moment. We define the calculations of our bot using algorithms. This is defined as a "...set of instructions designed to perform a specific task."⁶

Mathematically, a machine learning algorithm looks like this⁷:

³Expert System, *What is machine learning? A definition*, Expert System(2017), available at <https://expertsystem.com/machine-learning-definition/>.

⁴Unsupervised AI - AI for Complex Network Security - A Tutorial (2019). - <https://mixmode.ai/blog/whitepaper-unsupervised-ai-ai-for-complex-network-security/> Accessed 20 Oct 2020

⁵ibid

⁶Per Christensson, *Algorithm Definition*(2013), available at <https://techterms.com/definition/algorithm>.

⁷Jason Brownlee, *Master Machine Learning Algorithms - Discover How They Work And How To Implement Them From Scratch* (Jason Brownlee. 2019).

$$Y = f(X)$$

All this means is that some dependent variable Y (the answer for which we are looking) is equal to some function f (the formula we will use) multiplied by X (the independent variable that we are given). At least that is the way that we were taught in our high school algebra classes to read the equation. The “ f ”, of course, was some formula that we were expected to apply to X in order to calculate Y .

In machine learning, however, things are not quite the same as they were when we were in high school. In machine learning we are given neither X or Y . We don’t even start out by knowing f . That’s where training comes in.

Our training data is in the form of randomly selected X s and Y s that represent what we know and what we are seeking. We then do some calculations to discover what function f will give us those values.

For example, if we assume Y equals 4 and X equals 2, we immediately can see that f is equal to 2.

$$4 = 2(2)$$

If we keep f and try with $Y = 6$ and $X = 3$, we see that our f works out correctly.

$$6 = 2(3)$$

This is a greatly simplified example and in real life the X s, Y s and f would be much more complicated. We could use supervised training to input predetermined X s and Y s in order to figure out what f is. Then, we could see that whenever our bot sees the X it applies the f and calculates the Y .

In unsupervised training the bot is given nothing but the X s and it must derive f and calculate Y . It is a much slower but much more comprehensive process. The process can be sped up by using semi-supervised training and that gives us a start on deriving the f .

Because, these data are complicated, there may be no f that cleanly calculates Y from a particular X as in our examples above. Thus, we must change our algorithm slightly to accommodate that possibility,

$$Y = f(X) + e$$

e being the error. This error very likely is not a single number but is more likely another formula like f . In our examples above the error is zero so we ignored the e .

This is all the data science we need to begin to move into the legal realm and consider why the approach of adversarial machine learning is such a legal dilemma.

How the Autonomous Bot Likely Will Work

There are no truly autonomous bots yet (this writing – November 2019). But there is malware that in time will be candidates for conversion or upgrading to the ability to function autonomously.

Today these bots – or botnets as they sometimes are called – use a device called a command and control server (C2 or C&C). This server issues the bots their instructions, collects whatever data the C2 requires and communicates with its hacker human, usually called a “bot herder”.

The bots communicate periodically with the C2 based upon a predetermined schedule and algorithm that lets the bot find a C2 with which it can communicate. The bot herder communicates with the C2, collects stolen data and issues new instructions for the botnet.

The whole process is under the control of the bot herder and the C2 simply is an intermediary or “proxy” for the human bot herder. The goal of the autonomous bot is to eliminate the C2 and make the bot herder’s instructions superfluous.

Although there are no true autonomous bots yet, there are two developments that are very close and have, in fact, generated attacks: hivenets and swarmbots⁸. Although this may seem overly technical – and, perhaps, a little silly - in fact these two developments pose both a legal challenge and a legal opportunity.

Manky describes swarmbots as, “...a scalable architecture based on integrated and autonomous swarm intelligence.” This, actually, is far simpler than it sounds. Consider ants, for example. The ants work together to accomplish a task with each ant doing its particular job gaining insight and intelligence from the group – or swarm – of ants.

Star Trek fans will recall the Borg, a cybernetic collective with a “hive mind”. This refers to the sharing of a common mind, much as the ants do, and making decisions as a group based upon the knowledge of the individual members. A hivenet is much the same and the swarmbots are much the same as ants.

The benefit of a hivenet is that the swarmbots can reason collectively using data collected by the individuals. The hive learns through machine

⁸Derek Manky, *The Evolving Threat Landscape - Swarmbots, Hivenets, Automation in Malware*, CSO Online(2018), available at <https://www.csoonline.com/article/3301148/the-evolving-threat-landscape-swarmbots-hivenets-automation-in-malware.html>.

learning and once trained initially – perhaps using semi-supervised training to get a head start defining the parameters of the bot master’s intended mission. The swarm then can seek victims and gather its assigned data communicating with each other and, ultimately, the bot master – now the hive master - without using a C2.

Addressing Legal Challenges

From a legal perspective this poses significant challenge. For example, in what jurisdiction is the hivenet? From a technical perspective it could cover the globe, although that would be a large – but not impossible – hive. As with most cyber-related activities the most difficult issues are those involving jurisdiction.

Equally vexing, when a member of the swarm attacks a system – a banking network, for example – what forum has jurisdiction and, perhaps more importantly, over whom does the forum (of whatever location) have that jurisdiction? The member of the swarm that is attacking and exfiltrating data is under control of the hive, or, in *Star Trek* terms, “the Collective”.

The easy answer – and probably a correct one, although there are significant technical challenges involved - is the hive master or the programmer of the swarm.

There is another possible answer, though, and this is the one we favor: the jurisdiction of the victim over the hacker who gave the swarm its mission.

In his draft PhD dissertation⁹, Stephenson lays out a three-pronged approach for determining jurisdiction in cyberspace:

1. The Cyber Event Test, and...
2. The Modifier Test, and...
3. The Cyber Effects Test

⁹Peter Stephenson, A Framework for Determining U.S. Jurisdiction in Cyberspace (2019) University of Leicester, School of Law).

In order to be classified as a Cyber Event, the case must adhere to both the definition of cyberspace¹⁰ and the definition of cyber science¹¹ (the Cyber Event Test).

In order to be subject to jurisdiction in cyberspace, the case must demonstrate purposeful avilment within the context of cyberspace (the Modifier Test).

In order to be subject to jurisdiction in cyberspace the case must be able to apply the standard effects test within the context of cyberspace (the Cyber Effects Test).

Of the three prongs, the Modifier Test is the most potentially troublesome because demonstrating the bot herder's purposeful avilment of the victim's jurisdiction is arguable based upon a lack of specific target knowledge by the bot herder.

We begin addressing this challenge by understanding what is meant by "victim". The United States Code¹² gives us such a definition in the context of criminal law:

The term "crime victim" means a person directly and proximately harmed as a result of the commission of a Federal offense or an offense in the District of Columbia

The key phrase in this definition for our purposes is "proximately harmed". Proximate cause is tied to foreseeability. One may reasonably ask if the defendant should have foreseeably expected the results. Applying this to the Modifier Test one might ask further if the bot herder could foreseeably expect the victim to suffer an attack from the swarm that is managed by the hive that the hacker has programmed.

This does not require that the bot herder designate the victim explicitly. If the victim fits the general definition of the hive's target (a banking network or a particular weakness in a network's perimeter as examples), the bot herder should reasonably expected that the swarm – failing pro-

¹⁰Cyberspace is a complex global information infrastructure that facilitates communication between technology such as computers, networks and other digital systems, both independently and on behalf of people using it. Cyberspace per se is distinct from physical space and the constraints imposed by it such as geographic boundaries (Stephenson).

¹¹Cyber science is the study of phenomena caused or generated within the cyber space, which may or may not interact with phenomena caused or generated within the physical space (Stephenson).

¹²18 U.S. Code §3771 (e)(2)(A)

gramming errors, of course – could attack any target meeting the general definition, and, given enough time, conceivably would attack them all.

Technical Countermeasures and Their Implications in Law

We discussed the nature of machine learning above. Since machine learning can either be supervised or unsupervised, we now must note what technical countermeasures an adversary might use to circumvent machine learning in a system's defense mechanisms.

Technical Implications

These mechanisms focus upon learning the training set and calculating the function f . By calculating f , the adversary can determine how the defensive measures will respond to an attack and then craft an attack that circumvents those measures. This process – determining f and extracting the training set is called *adversarial machine learning* (AML).

The literature on AML is voluminous – nearly 1500 papers since 2014¹³. The key issues in an AML attack are gaining knowledge of the target and its defenses and deducing the classifier's algorithm. Performing this type of “learning” attack requires interaction with the target and interaction with the target implies the ability to ferret out the attacker. This, obviously, is not desirable on the part of the attacker.

Should the attacker be deduced, the three-prong test discussed in section III above would suffice for determining jurisdiction. However, thanks to botnets – and, more specifically, autonomous botnets or hives – the swarm can be programmed to perform the analysis and inject a successful attack. Such an attack is DeepFool, an “... algorithm to efficiently compute training data perturbations that fool deep networks...”¹⁴

While the notion of autonomous botnets – hives of swarmbots – is worrying, the ability of the swarm to derive countermeasures to an AI defense are even more troubling. The technology to analyze a defensive system, calculate f , deduce the classifier and create a successful adversar-

¹³Nicholas Carlini, *A Complete list of all (arXiv) Adversarial Example Papers*, Nicholas Carlini(2019), available at <https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html>.

¹⁴Alhussein Fawzi Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, *DeepFool: a simple and accurate method to fool deep neural networks*, (2015).

ial training set¹⁵ is beyond the capability of any practical single bot. But using the combined intelligence of the swarm makes such an attack feasible.

Legal Implications

Obviously, if we are dealing with a hive of swarmbots – the hive, and, thus, the bots operating autonomously - pinpointing jurisdiction is potentially very difficult. Applying our three-prong test we find:

The Cyber Event Test

This clearly is a cyber event. In fact, we may consider it a prototypical cyber event in that it occurs completely in cyberspace with virtually no human interaction (beyond the original programming and harvesting of the results of the attack, of course).

The Modifier Test

Certainly, the modifier test applies. The hive, and, arguably, elements of the swarm have purposely availed themselves of access to the target. What that availment means is relative to the type of attack and the vulnerabilities exhibited by the target. In addition to purposeful availment, minimum contacts also may be demonstrated¹⁶.

The Cyber Effects Test

Equally, there is no question about the impact of the attack. The target is the place of the harm.

¹⁵Igino Corona Battista Biggio, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, Fabio Roli, *Evasion Attacks against Machine Learning at Test Time*, Lecture Notes in Computer Science 387(2013).

¹⁶Even if only a single contact such as a probe. See *McGee v. International Life Insurance Co.*, 355 U.S. 220, 78 S.Ct. 199, 2 L.Ed.2d 223 (1957) – nature and quality of the contact(s).

The Law V. Technology in Complicated Applications of Personal Jurisdiction

The approach to determining jurisdiction in common use in the United States is well-demonstrated in a case brought by Microsoft¹⁷. In this case Microsoft identified 82 command and control servers around the world and attempted to serve their system administrators.

Of course, none responded so Microsoft simply obtained a cease and desist order for each server operator. They then proceeded to take control of the servers and disable them under authority of the order. While that case has its own difficulties, it is emblematic of typical botnet cases in America today.

However, in a world of autonomous hives and swarmbots investigators and legal authorities do not have the luxury of an identifiable command and control server because there is none. However, the hive must exist at some place in cyberspace.

The primary difference between a hive/swarmbots architecture and a command and control architecture is that with a command and control architecture the bots are in communication with the C&C server which, in turn, is in communication with the bot herder. In a hive, the hive launches bots and both the bots, and the hive operate autonomously without recourse to a bot herder or bot master. While there are other, currently in use, mechanisms that do not involve AI, the direction of fully autonomous control represents the next wave of malware distribution technology simply because it is autonomous.

This leaves us with two vexing challenges; locating the hive and establishing jurisdiction. We will not address locating the hive because that is a challenge for technologists. For our purposes here we will assume that the hive has, in some form or manner, been located.

The swarmbots communicate and exchange data with each other and the hive, presumably, simply provides program changes to the swarm. This could be by transmitting the change to a single swarmbots which disseminates the change throughout the rest of the swarm and then destroys itself to prevent traceback to the hive.

Likewise, the operator of the hive may destroy it after placing a copy in another location. All communications are outbound to the hive and outbound from the hive to the swarm. This prevents the possibility of “capturing” a bot, reverse engineering it, and determining where the hive is. It simply does not know, unlike a C&C architecture. More important, the hive can self-learn and join with other hives. Demonstration hivenets

¹⁷Microsoft Corp. v. John Does 1-82, No. 3:13-CV-00319-GCM, 2013 WL 6119242 (W.D.N.C. Nov. 21, 2013)

with nearly 3 billion botnet communications in a single quarter have been demonstrated in 2017¹⁸.

This leaves the knotty legal question, “Who owns the hive?” The hive likely will be located in one two places; a server owned by the operator (unlikely) or a server owned by another entity and compromised by the hive operator. As the hives self-learn and launch more swarmbots, the number of hives will increase, again, without assistance of a human operator.

Addressing the compromised server (“intermediate server”) alternative, we are faced with the question of who does the victim look to for the purpose of establishing jurisdiction over the bot that compromised his (the victim’s) system. Stephenson is clear that an owner of a compromised server may not be the target of a legal action but may initiate legal action against the entity that compromised his/her server¹⁹.

This leaves us with a dead-end unless technical experts can identify the attacker of the victim server being used to house one or more hives. That certainly can pose difficult challenges if the intermediate server is not in a friendly jurisdiction.

One approach, suggested by the Microsoft case, is identification of the true owner of the hive by tracking Internet Protocol (IP) registration. An IP is the “address” of a system on the Internet and is assigned when the new system goes online. However, there are other difficulties including the use of VPNs (virtual private networks), unregistered Ips and use of either the TOR network or other, perhaps private, darknets. From a legal perspective this is fraught with problems.

The first issue to be wary of is the ability to identify the registrant of the IP address. It is typical of IP registrants that they do two things. First, they hide behind an Internet alias. Aliases are very easy to create and are extraordinarily realistic.

For example, we created an alias named Diana Schmidt with the following characteristics²⁰:

¹⁸Derek Manky, Fortinet Predicts Highly Destructive and Self-learning “Swarm” Cyberattacks in 2018 (Fortinet 2017).

¹⁹Peter Stephenson, *International Private Law as a Model for Private Law Jurisdiction in Cyberspace*, 7 Legal Issues Journal (2019). “**Guideline 3** – Non-consensual or unauthorized control of Internet or other network-attached devices and computers does not dictate personal jurisdiction in the forum where the devices or computers reside.”

²⁰Fake Name Generator, *Your Randomly Generated Identity*. Fake Name generator online tool (2019), available at <https://www.fakenamegenerator.com/gen-random-gr-gr.php>. The tool is made by Coraban Works which appears to be in San Antonio, Texas, an apparently one-person shop owned by Jacob Allred which may or may not be the owner’s real name or location.

- Nationality: German
- Address: Alter Wall XX, Lohr, Germany (with geo coordinates)
- Mother’s maiden name: Pfeiffer
- Phone: (actual, though not active, with country code)
- Birthday: 31 Aug 1974
- Age: 45
- Email address: DianaSchmidt@jourXXXXXX.[com]
- Username: (given but redacted here)
- Password: (given but redacted here – both username and password work)
- Mastercard: (real pattern but not usable to make purchases)
- Full employment details
- Height, Weight and Blood Type
- UPS, Western Union and MoneyGram tracking numbers
- Vehicle details

This tool is free to use, readily available on the Internet and users may select from dozens of nationalities and countries. Once an Internet alias is created, a “handle” or nickname comes next. Typically, the actor will not use the alias’ email but will establish an anonymous, encrypted email account using a free service such as ProtonMail²¹. For that account, the actor will use his or her handle. Using that account, the actor will register the IP address. That means that even if the actor is traced back through his or her handle to the alias – in this case, Diana Schmidt – the trail will end there. Service of subpoenas to the alias is likely to fail.

Additionally, these actors often use a service to mask their registration details. This is called “whois protection”. Whois is the open registration database for IP addresses on the Internet. By using whois privacy the only thing an investigator sees upon conducting a domain search is the information for the privacy provider. Figure 1 shows an example²².

²¹<https://mail.protonmail.com/login>

²²name.com, *Get Advanced Security for Your Domains*, name.com(2019), available at <https://www.name.com/whois-privacy>.

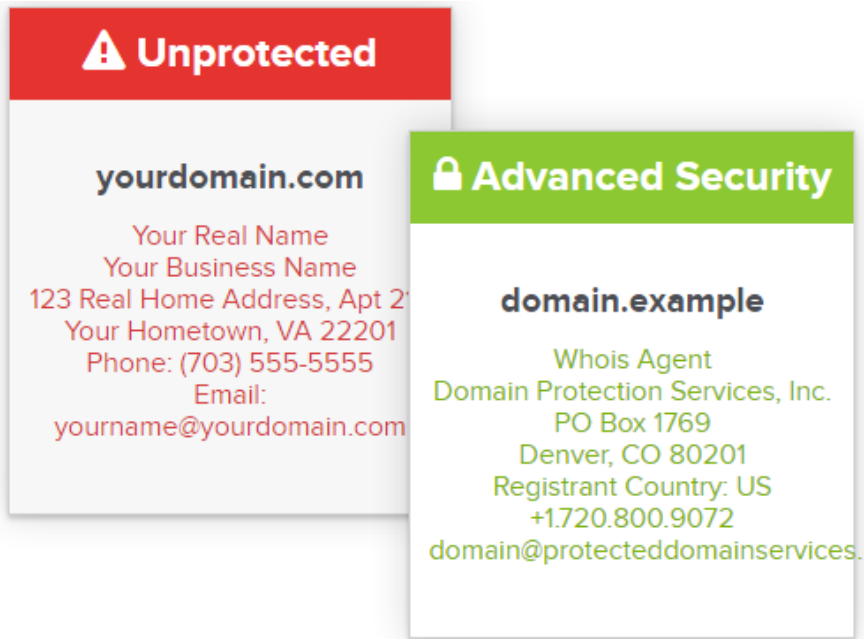


Figure 1. Whois privacy as provided by name.com

The challenge from a technical perspective is clear: identifying the actual owner of a hive IP. Returning to the Microsoft case, the same challenge exists when identifying the bot herder of a current generation botnet.

Microsoft addressed the problem simply by attempting to serve the 81 domain registrants at their individual domain name service providers. All attempts, of course, were unsuccessful and the court determined that they had been constructively served.

Because personal jurisdiction can be waived, as happens often when the service is unsuccessful, the Microsoft court made the assumption of constructive service and granted the plaintiff's plea for cease and desist orders as well as the takedown of the servers.

He problem, of course, is that should one or more defendants respond to the subpoena and should those defendants not be in the jurisdiction of the forum it is questionable if personal jurisdiction would have been granted.

When dealing with hivenets, however, the issue becomes more complicated. It will be extremely difficult if not – given the then state of the art – impossible to identify the location in cyberspace where the hives reside. That is exacerbated by the hive's ability to reproduce itself without human intervention. It can be programed to reproduce, seek a new home and register itself using the obfuscation techniques (at minimum) we have discussed.

Thus, a single hive can be launched and programmed to reproduce, register and, after some predetermined instances destroy itself and let it progeny continue creating new hives and launching swarmbots. In that case there would be little or no trail and little or no way to identify the original hive operator. The likely approach – at least as we understand hives and swarms at present – would be to seek out the locations where the swarmbots deliver their booty after a successful breach of a victim.

These are technical issues, to be sure, but they inform the legal solutions available to establish personal jurisdiction over the original hive developer. Another legal challenge assumes that the original developer sells or rents hives and swarms to other actors. This is common practice today with botnets. The question then becomes, “whom does one sue?”

There is very little precedent in American law relating to this problem. Virtually all such cases in Westlaw relating to botnets are dominated by actions initiated by Microsoft. All these cases are against John Does since Microsoft developed no direct knowledge of the identities of the actors.

There are, however, a few cases that do not involve Microsoft. In these cases, the actor is known²³. Where the actor – in these cases the bot herder or bot master – is known standard rules for determining personal jurisdiction apply. In cases such as the Microsoft “Doe” cases determining whom to serve is more problematic.

Returning to Adversarial Machine Learning and the Legal Challenge

One would expect that a target equipped with artificial intelligence/machine learning defensive tools would be safe from the impact of an attack by swarmbots. After all, the defensive measures are of a type with the adversarial measures of the swarm. In the face of adversarial machine learning, however, that may not be the case.

While we have discussed the difficulty of establishing the identity of the creator of a specific hive. We have yet to address what happens when the attack occurs and succeeds. If the attack occurs, and fails after penetration of the target, it is conceivable that investigators will be able to extract samples of the swarm bots for analysis.

²³See, for example, 729 Fed.Appx. 112 (Mem) United States v. Gasperini, 729 F. App'x 112 (2d Cir. 2018), United States v. Hussain, No. 16-CR-00462-CRB-1, 2017 WL 4865562 (N.D. Cal. Oct. 27, 2017) and United States v. Golightley, No. 18-10097-JWB, 2019 WL 1458949 (D. Kan. Apr. 2, 2019)

If the attack succeeds, however, such samples may not exist. The swarm bot may be instructed in its programming shared with the hive(s) to deliver its purloined payload out of the victim system, obfuscate the exfiltration and destroy itself removing all evidence.

So, the operative question becomes, “how would the attack succeed in the face of ML-equipped defenses?” The answer may lie in the adversary’s application of AML.

For adversarial machine learning to occur, the adversary must do several things. What those things are depends upon the adversary’s chosen attack model. There are three²⁴. They all need to retrain the classifier to recognize a malicious training element (a “perturbation”) as non-malicious. It is the classifier’s job to classify elements as malicious or not and it uses an algorithm for that purpose.

White box attacks – the attacker knows everything about the target’s model and can reproduce it, making small perturbations in it to fool the classifier and retraining the victim. This is almost always likely to be an insider attack since access to the underlying algorithm and the original training set are necessary.

Gray box attacks – The attacker knows the underlying algorithm, the model topology and nothing else. The training set either must be deduced or captured in some way before the classifier may be reprogrammed to believe that introduced perturbations in the training set are real.

Black box attacks – The attacker knows nothing about the target and must derive the training set and the underlying algorithm for the classifier. The attacker does this by “querying the oracle”. That consists of getting the target to respond to a large variety of attacks and collecting the results. Having done that, the attacker may be able to reproduce the training set and the classifier algorithm allowing retraining of the target.

All AML attacks fall into one of these three categories. The ability of the law to establish personal jurisdiction depends upon what data about the hive master that the technical investigative team can uncover from the residue of the attack.

The biggest legal challenge is availability of data. Without identifying data, the Microsoft approach discussed above is our best bet and, as we showed, not particularly reliable. What is needed is a closer partnership between lawyers and cyber technologists. Beyond that, a new area of cyber forensic science is necessary.

Almost no universities teach malware forensics. In a sample of the top 34 universities worldwide teaching computer science only 14 teach malware analysis, the starting point for the level of cyber forensics needed

²⁴Mohammad Samragh Bitar Darvish Rouani, Tara Javidi, Farinaz Koushanfar, *Safe Machine Learning and Defeating Adversarial Attacks*, 17 IEEE Security & Privacy 31(2019).

here²⁵. While malware analysis is a very specialized field, malware forensics is even more so. Clearly, universities have yet to see the benefits of addressing such arcane knowledge. Those that do, however, are among the preeminent educational institutions teaching technology in the world.

The most difficult problem in identifying a cyber adversary is attribution. Lin emphasizes the difficulty of attribution, its many faces and approaches to solving the problem²⁶ by citing Rid and Buchanan²⁷ "... thinking about attribution is currently based on three assumptions ...". Those assumptions are,

1. "attribution is a largely intractable problem because of the technical characteristics and the geography of the Internet"
2. "attribution is either possible or not possible in any given case of interest"
3. "the main challenge in attribution is *finding the evidence itself* and not in interpreting or using it" (emphasis added)

All three assumptions, true in 2014 are equally true now. The only difference is that there now is a fourth assumption that we add here as a corollary to the third:

4. The main challenge in *finding the evidence* may be attributed to the increasing complexity of malware including botnets, swarmbots, hivenets and machine learning, including adversarial machine learning.

Thus, we have a circular problem. First, we have the technical characteristics of the Internet. Second, we have the challenge of finding the evidence. Third, we have the prickly issue of jurisdiction in cyberspace, in part due to the technical characteristics of the Internet.

Stephenson takes up the third issue – jurisdiction in cyberspace – and addresses it with the three-prong test²⁸. But Rid and Buchanan's third assumption is, we submit, really the first and most difficult to overcome. It is difficult because it subsumes assumptions one and two.

²⁵Swetha Gorugantu, *Malware Analysis Skills Taught in University Courses* (2018) Wright State University, The Graduate School).

²⁶Attribution of Malicious Cyber Incidents. No. Aegis Series Paper No. 1607(2016).

²⁷Thomas Rid and Ben Buchanan, *Attributing Cyber Attacks*, 38 *Journal of Strategic Studies* (2014).

²⁸See Section III above

Taken together, if one can solve the challenges leading to the first three assumptions and the corollary number 4, one can manage jurisdictional conflicts in any number of standard, well-established ways.

The legal community must, therefore, address its forensic needs with the technical community so that the technical community can provide *evidence*, not, simply, data. The evidence that the legal community needs is focused upon the requirements for attribution as laid out by Lin in his Aegis Series paper (see footnote citation).

Exploring a Hypothetical

An easy way to explain the issues involved is the use of a hypothetical attack. To our knowledge, no attack such as this has yet occurred. However, it is entirely plausible, and all the technologies – if not the products – employed here currently exist.

In this attack, the victim is a bank located in San Francisco. San Francisco is in the American “Silicon Valley” and is one of the hearts of American cyber technology know-how and education.

This bank has highly developed cyber defense systems including such capabilities as artificial intelligence-based monitoring. The monitoring uses unsupervised training algorithms for its machine learning functionality. It is capable of detecting, classifying and responding to imminent threats as soon as the threats hit its perimeter. It is constantly learning from the network traffic. From the outside (Internet), anyway, it appears impenetrable.

Over a weekend the on-duty help desk engineer notices that several accounts are being accessed simultaneously and their contents exfiltrated. This amounts to a mass withdrawal of money and payment card information from all accounts – saving, checking, etc. – on the affected customers.

However, analysis of the readouts of the cyber defense systems show nothing wrong. There appears to be no attack. Clearly something is amiss however because over 500 accounts are being simultaneously emptied. There is no indication of where the attack is coming from and no indication of where the money is being exfiltrated to. The help desk engineer severs the bank’s Internet connection and calls for help.

The forensic engineers respond immediately and begin to perform cyber forensic analysis on the affected servers to determine a cause. They find no evidence of malware. Whatever bot was in the system has erased its tracks and destroyed itself. Loss to the bank’s 500 affected customers is over a million dollars. The attack appears to have taken less than 5 minutes to succeed and has left no trace, even within the sophisticated AI defenses.

The bank's AI defense systems are, as we pointed out, unsupervised learning systems. That means that it develops its ML training set "on the fly" and there is no preconceived training set. The benefit to that is that the unsupervised training set cannot be guessed as easily as a preconceived training set used in supervised learning. In other words, a supervised learning training set is just a database that can be stolen, modified and returned to reprogram the ML.

Conversely, an unsupervised training set notices everything that happens on the network and adds it to its database. Were it to see an effort to reprogram it, the AI would interdict the effort. The only way to do an attack is querying the oracle²⁹, recreating the training set empirically and attempting to replace portions of the existing training set with the modified one. Attacking the training set directly both would not work and any bot entering the network for that purpose would be intercepted.

Obviously, it develops that the only way this black box attack could have succeeded would have been querying the oracle, creating a nearly duplicate training set and imposing portions of the training set on the target. The easiest way to do that is to take the supposed training set created by querying the oracle, creating a small perturbation that would cause no damage and repeatedly hit the target with it using a large number of swarmbots. Since it causes no damage the ML decides it is OK – normal operation – and starts to ignore it.

Once the ML starts to ignore it, the attacker might introduce a bot that looks to the ML as harmless and that bot starts exfiltrating user credentials. The exfiltration might look to the ML like legitimate web site accesses. Once the bot is finished it cleans up all traces of its activity and destroys itself. There is no indication that it ever was in the bank's system and all the cyber forensics engineers in Silicon Valley cannot find its artefacts.

Now, on the weekend when the attacker is sure that the bank is on a skeleton staff it performs what security professionals refer to as a "smash-and-grab". At a predetermined time, using a preprogrammed script, swarmbots impersonating all bank customers whose credentials were stolen "log in" to their accounts. The system does not see this as an attack because proper credentials are being used and no withdrawal policies are being violated.

The accounts are emptied to remote servers that immediately ship the money to anonymous Bitcoin accounts and the servers destroy themselves and any evidence of their existence. The attack has succeeded completely thus far, cyber forensic engineers have nothing and the bank's legal team – even with the assistance of outside cyber lawyers – has no evidence to pursue.

²⁹See Section V, Black Box Attacks

We could make this more difficult still to track if we specify that all authentication and identification of and to the servers that receive the money payloads be created using the blockchain³⁰. This creates the ultimate (today) anonymity. That means that the only place where a trail might be picked up – the exfiltration – is obfuscated by the blockchain. This constitutes an almost perfect next generation attack.

We now must ask what resources the adversary would need to accomplish this attack successfully. The use of a hivenet with thousands of swarmbots would accomplish the attack nicely. The original hive is programmed to accomplish the mission of emptying a million dollars from the bank's customer accounts. The hive then creates the hivenet without human interventions and destroys itself destroying any connection to its creator. The hivenet creates the swarmbots necessary to accomplish the mission, again, without human intervention. The attack could be coordinated to run simultaneously against several banks across the globe with disastrous results.

Today there are no laws that are designed to address this. If we could gather the evidence we could, perhaps, establish jurisdiction in San Francisco. But jurisdiction over who? We don't even have enough evidence pointing to IP addresses and their supposed owners to take the Microsoft approach.

In United States law you cannot establish personal jurisdiction if you cannot serve the defendant. Microsoft assumed that the IP addresses were registered by their owners (even though the aliases made that unlikely) and served the pseudo owners.

Here we have no owners of IP addresses because the IP addresses were created using blockchain technology³¹ and all communication is encrypted and via the blockchain. The attack succeeds and there is no forensic evidence for the lawyers to use. Were there such evidence the legal issues would simplify greatly.

This hypothetical illustrates the need for close collaboration between legal and cyber experts. It also illustrates the need for more and better-trained cyber lawyers and cyber experts.

³⁰A blockchain is a database of "blocks" linked cryptographically. Only the cryptographic hash can access the contents of the block and there is no non-blockchain identification associated with it. It is used commonly by cryptocurrencies such as Bitcoin but has numerous other applications as well. – various sources including www.medium.com/@Cardanians_io/a-database-is-not-a-blockchain-7027a319311e accessed 20 Oct 2020

³¹See <https://blockchain-dns.info/> with the understanding that the technology likely would be used but no service such as this one would offer the requisite anonymity.

Conclusions and Possible Solutions

For as dire as the hypothetical sounds, there are solutions. However, the solutions are not simple, and they are global. The finding of the very small number of universities globally that teach the skills needed to start a career in malware forensics is quite troubling.

An Emerging Role of Cyber-Legal Practitioner Is Needed

Further, a global recognition of what the field of cyber-legal practice requires on the technical side. The American Bar Association is typical of key legal organizations underplaying both the role of a cyber-legal practitioner and underplaying the educational requirements. It takes the position that³², "...you do not need to major in computer science, information technology, or cybersecurity—though certainly that can help." The author does not agree.

Consider the medico-legal field. Most medico-legal practitioners have an advanced degree in both the law and medicine – usually a JD and an MD. Our position is that cyber-legal practitioners have no less need for advanced education. Certainly, a JD (if one wishes to practice) or PhD/LLD (if one wishes only to advise) in law and a PhD in cyber science, cyber security or a related field are necessary.

This presumes that there is an actual field called cyber-legal practice. A brief survey of the Internet using Google and searching for "+cyber +legal +practice" revealed that the common definition is nothing close to the issues discussed in this paper. The focus is on cyber law and addresses the usual public and private legal issues in an Internet context.

However, like the medico-legal field, the cyber-legal field is fraught with special problems, most arising from forensic evidence in the emerging autonomous world of the Internet. Consider, for example, a defamation case in which an individual is defamed in a public forum such as Facebook or Twitter.

If the defamer is an individual who can be identified the usual findings around jurisdiction and the elements of defamation are met. The suit goes to court. However, suppose the defamer is an autonomous bot. Today we know that such bots routinely make political statements in online forums, it is no stretch at all that such bots – not even autonomous – could do likewise and defame.

³²Leonard Wills, *How to Become a Cybersecurity Lawyer*, American Bar Association(2018), available at <https://www.americanbar.org/groups/litigation/committees/minority-trial-lawyer/practice/2018/how-to-become-a-cybersecurity-lawyer/>.

Consider some of the technologies discussed in our hypothetical. Clearly such bots – now part of a swarm with instructions from a hivenet - could do the same thing in online forums. Today’s “cyber lawyer” would be helpless without outside help.

Cyber scientists, likewise, would not have the legal insights to understand what evidence would need to be discovered without legal training. The role of the cyber-legal specialist is to know both sides of the equation and help bring the defamation to a satisfactory conclusion.

Cyber Forensic Training Including Advanced Malware Analysis and Evidence Identification and Gathering Needs To Enter Education Programs

As we mentioned earlier in this paper and last year³³ Digital – or “cyber” if one prefers – forensics education is in a sad state. In 2018 our position was that even the designation “digital forensic science” was arguable as to its accuracy. That may or may not still be the case but in either case it is of less concern here than the depth of the education required to make such a discipline useful for the legal community.

The legal community depends for its success on evidence. One can testify ad nauseum but if evidence does not support those claims the case will not go well. The problem, as we have abundantly illustrated, is significantly exacerbated when the stakes are high, and the adversary takes advantage of technologies such as artificial intelligence.

The solution is a difficult one both from the perspective of willingness of universities to add courses that are not profitable – no matter how much they benefit society – and from the more important perspective of qualified instructors. The solution to this challenge truly is a “bootstrap” operation.

Lawyers and Courts Need To Become Better Acquainted With Advanced Cyber Threats

While there may be ample income in cyber law as it is seen today (which we doubt given some interviews with cyber lawyers) there are very few law firms that view cyber law as a practice in itself. We contacted several self-designated cyber lawyers in the preparation of this paper and found that they put the designation on their web sites for marketing purposes, but they really have no practice in the field. What they do have generally

³³P. R. Stephenson, *Digital Forensic Science: An Oxymoron?*, 6 Legal Issues Journal 95(2018).

is built around child pornography, defamation and infringement of intellectual property. None of these are inherently cyber crimes even though they may have cyber elements^{34, 35, 36}.

We have been privileged to work with one very large law firm that had conducted an intricate cyber investigation of a complicated (at the time) breach. We were called upon to validate the results of a private security firm's investigation of the breach.

The lawyers we worked with had an excellent knowledge of the technology involved and we ended up coming to the same conclusion. Although the primary lawyer we worked with did not have a PhD in cyber science he was a prototype of the type of cyber lawyer that our society needs today.

Conclusions

There are some important conclusions to draw from this paper and the research behind it.

First, the stage on which malicious cyber event play out is changing dramatically. Although well over 80% of so-called cyber-related cases

³⁴The web site "Legal Career Paths" defines cyber law as "...the area of law that regulates how people use the internet" and lists several different types including, fraud, copyright, defamation, harassment and stalking, freedom of speech, trade secrets, and contracts and employment law. It mentions nothing that resembles our hypothetical. <https://legalcareerpath.com/what-is-cyber-law/> accessed 20 October 2020

³⁵The American Bar Association in its "Techreport 2018" published in January of 2019 states: "Attorneys also have common law duties to protect client information and often have contractual and regulatory obligations to protect information relating to clients and other personally identifiable information, like health and financial information. These duties present a challenge to attorneys using technology because most are not technologists and often lack training and experience in security. https://www.americanbar.org/groups/law_practice/publications/techreport/ABATECHREPORT2018/2018Cybersecurity/ accessed 20 October 2020.

³⁶E. F. G. Ajayl, *Challenges to enforcement of cyber-crimes laws and policy*, 6 (1) Journal of Internet and Information Systems (2016). - "Related to the above factors of poor training, remuneration and inadequate security and protection on the hazardous job for law enforcement agency officials is the dearth of experts in the prosecution of cybercrimes."

can be solved without recourse to cyber technology³⁷, there is an increasing number that are hard cases³⁸ from the cyber perspective.

For those cases that involve artificial intelligence in general and machine learning in particular, the challenges border on the extreme as our hypothetical illustrates. We are not, today, prepared legally or technically to address these new challenges. However, and much more important, lawyers and cyber subject matter experts (SMEs) are not yet prepared educationally to work together.

When the law takes on the creator of an autonomous malware system (hivenets and swarmbots, for example) it will have to depend upon expert witnesses and even then the interactions are very complicated. Our conclusion is that the better solution is to develop a cadre of cyber-legal expert practitioners who can aid the legal system in responding to the enhanced cyber threats that we see on the horizon. Perhaps even more important, though, is the obvious prediction that nation-states and sub-state actors will be among the early adopters of these malicious technologies due to their resources and global objectives. That suggests that the cyber-legal aspects of artificial intelligence need to become part of every nation's cyber governance.

We are not there yet and the clock is running out as more and more AI appears in virtually every aspect of our society, especially cyber security.

Bibliography

- Expert System, *What is machine learning? A definition*, Expert System(2017), available at <https://expertsystem.com/machine-learning-definition/>.
- Unsupervised AI - AI for Complex Network Security - A Tutorial (2019). Per Christensson, *Algorithm Definition*(2013), available at <https://techterms.com/definition/algorithm>.
- Jason Brownlee, *Master Machine Learning Algorithms - Discover How They Work and How to Implement Them From Scratch* (Jason Brownlee. 2019).
- Derek Manky, *The Evolving Threat Landscape - Swarmbots, Hivenets, Automation in Malware*, CSO Online(2018), available at <https://>

³⁷Research for the author's as yet unpublished law PhD dissertation on cyber jurisdiction

³⁸Martin Stone, *Formalism*, in *Oxford Handbook of Jurisprudence and Philosophy of Law* (Jules Coleman; Scott Shapiro ed. 2004). The author distinguishes between hard cases and easy cases, easy cases being those for which a solution is obvious using well settled law

- www.csoononline.com/article/3301148/the-evolving-threat-landscape-swarmbots-hivenets-automation-in-malware.html.
- Peter Stephenson, *A Framework for Determining U.S. Jurisdiction in Cyberspace* (2019) University of Leicester, School of Law).
- Nicholas Carlini, *A Complete list of all (arXiv) Adversarial Example Papers*, Nicholas Carlini(2019), available at <https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html>.
- Alhussein Fawzi Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, *DeepFool: a simple and accurate method to fool deep neural networks*, (2015).
- Igino Corona Battista Biggio, Davide Maiorca, Blaine Nelson, Nedim Srdic, Pavel Laskov, Giorgio Giacinto, Fabio Roli, *Evasion Attacks against Machine Learning at Test Time*, Lecture Notes in Computer Science 387(2013).
- Derek Manky, Fortinet Predicts Highly Destructive and Self-learning “Swarm” Cyberattacks in 2018 (Fortinet 2017).
- Peter Stephenson, *International Private Law as a Model for Private Law Jurisdiction in Cyberspace*, 7 Legal Issues Journal (2019).
- Fake Name Generator, *Your Randomly Generated Identity*. Fake Name generator online tool (2019), available at <https://www.fake-namegenerator.com/gen-random-gr-gr.php>.
- name.com, *Get Advanced Security for Your Domains*, name.com(2019), available at <https://www.name.com/whois-privacy>.
- Mohammad Samragh Bitar Darvish Rouani, Tara Javidi, Farinaz Koushanfar, *Safe Machine Learning and Defeating Adversarial Attacks*, 17 IEEE Security & Privacy 31(2019).
- Swetha Gorugantu, *Malware Analysis Skills Taught in University Courses* (2018) Wright State University, The Graduate School).
- Attribution of Malicious Cyber Incidents. No. Aegis Series Paper No. 1607(2016).
- Thomas Rid and Ben Buchanan, *Attributing Cyber Attacks*, 38 Journal of Strategic Studies (2014).
- Leonard Wills, *How to Become a Cybersecurity Lawyer*, American Bar Association(2018), available at <https://www.americanbar.org/groups/litigation/committees/minority-trial-lawyer/practice/2018/how-to-become-a-cybersecurity-lawyer/>.
- P. R. Stephenson, *Digital Forensic Science: An Oxymoron?*, 6 Legal Issues Journal 95(2018).
- E. F. G. Ajayl, *Challenges to enforcement of cyber-crimes laws and policy*, 6 (1) Journal of Internet and Information Systems (2016).
- Martin Stone, *Formalism*, in Oxford Handbook of Jurisprudence and Philosophy of Law (Jules Coleman; Scott Shapiro ed. 2004).