Active Cyber Defense: Applying Air Defense to the Cyber Domain[1]

Dorothy E. Denning & Bradley J. Strawser
Naval Postgraduate School

In the domain of cyber defense, the concept of active defense is often taken to mean aggressive actions against the source of an attack. It is given such names as "attack back" and "hack back" and equated with offensive cyber strikes. It is considered dangerous and potentially harmful, in part because the apparent source of an attack may be an innocent party whose computer has been compromised and exploited by the attacker.

Our purpose in writing this paper is to show that active cyber defense is a much richer concept that, when properly understood, is neither offensive nor necessarily dangerous. Our approach is to draw on concepts and examples from air defense to define and analyze cyber defenses. We show that many common cyber defenses, such as intrusion prevention, have active elements, and we examine two case studies that employed active defenses effectively and without harming innocent parties. We examine the ethics of active cyber defenses along four dimensions: scope of effects, degree of cooperation, types of effects, and degree of automation. Throughout, we use analogies from air defense to shed light on the nature of cyber defense and demonstrate that active cyber defense is properly understood as a legitimate form of defense that can be executed according to well-established ethical principles.

We are by no means the first authors to address the ethics of active defense. Dittrich and Himma (2005), for example, contributed substantially to initial thinking in this area. Our work differs from theirs and other work in this area through its application of air defense principles. We believe that the analogy of air defense helps shed light on active cyber defense and the moral issues it raises.

DEFINING ACTIVE AND PASSIVE CYBER DEFENSE

Because our definitions of active and passive cyber defense are derived from those for air defense, we begin by reviewing active and passive air and missile defense.

Active and Passive Air and Missile Defense

Joint Publication 3-01, *Countering Air and Missile Threats*, defines active air and missile defense (AMD) as: "direct defensive action taken to destroy, nullify, or reduce the effectiveness of air and missile threats against friendly forces and assets." The definition goes on to say that active AMD "includes the use of aircraft, AD [air defense] weapons, missile defense weapons, electronic warfare (EW), multiple sensors, and other available weapons/capabilities." (JP 3-01 2012) Active AMD describes such actions as shooting down or diverting incoming missiles and jamming hostile radar or communications.

An example of an active air and missile defense system is the Patriot surface-to-air missile system, which uses an advanced aerial interceptor missile and high performance radar system to detect and shoot down hostile aircraft and tactical ballistic missiles (Patriot 2012). Patriots were first deployed in Operation Desert Storm in 1991 to counter Iraqi Scud missiles. Israel's Iron Dome anti-rocket interceptor system has a similar objective of defending against incoming air threats. According to reports, the system intercepted more than 300 rockets fired by Hamas from Gaza into Israel during the November 2012 conflict, with a success rate of 80 to 90 percent (Kershner 2012). At the time, Israel was also under cyber assault, and Prime Minister Benjamin Netanyahu said that the country needed to develop a cyber defense system similar to Iron Dome (Ackerman and Ramadan 2012).

Another example of an active air defense system is the U.S.'s Operation Noble Eagle (Air Force 2012). Launched minutes after the first aircraft was hijacked the morning of September 11, 2001, the operation has become a major element of

homeland air defense that includes combat air patrols, air cover support for special events, and sorties in response to possible air threats. Although Noble Eagle pilots can potentially shoot down hostile aircraft, so far none have done so. However, they have intercepted and escorted numerous planes to airfields over the years.

In contrast to active defense, passive air and missile defense is defined as: "all measures, other than active AMD, taken to minimize the effectiveness of hostile air and missile threats against friendly forces and assets," noting that "these measures include detection, warning, camouflage, concealment, deception, dispersion, and the use of protective construction. Passive AMD improves survivability by reducing the likelihood of detection and targeting of friendly assets and thereby minimizing the potential effects of adversary reconnaissance, surveillance, and attack." (JP 3-01 2012) Passive AMD includes such actions as concealing aircraft with stealth technology. It covers monitoring the airspace for adversary aircraft and missiles, but not actions that destroy or divert them.

Active and Passive Cyber Defense

We adapt the definitions of active and passive air defense to the cyber domain by replacing the term "air and missile" with "cyber." This gives us the basic definitions: active cyber defense is direct defensive action taken to destroy, nullify, or reduce the effectiveness of cyber threats against friendly forces and assets. Passive cyber defense is all measures, other than active cyber defense, taken to minimize the effectiveness of cyber threats against friendly forces and assets. Put another way, active defenses are direct actions taken against specific threats, while passive defenses focus more on protecting cyber assets from a variety of possible threats.

Using these definitions, we now examine various cyber defenses to see whether they are active or passive. We begin with encryption, which is clearly a passive defense. It is designed to ensure that information is effectively inaccessible to adversaries that intercept encrypted communications or download encrypted files, but takes no action to prevent such interceptions or downloads. Steganography is similarly passive. By hiding the very existence of information within a cover such as a photo, it serves as a form of camouflage in the cyber domain. Other passive defenses include security engineering, configuration monitoring and management, vulnerability assessment and mitigation, application white listing, limiting administrator access, logging, backup and recovery of lost data, and education and training of users. None of these involve direct actions against a hostile threat.

User authentication mechanisms can be active or passive. For example, consider a login mechanism based on usernames and passwords that denies access when either the username or password fails to match a registered user. We consider this passive if no further action is taken against an adversary attempting to gain access by this means. Indeed, the person might try again and again, perhaps eventually succeeding. Now suppose that the mechanism locks the account after three tries. Then it has an active element in that this particular adversary will be unable to gain entry through that account, at least temporarily. However, it does not stop the adversary from trying other accounts or trying to gain access through other means such as a malware attack. Nor does it prevent an attacker who stole an account and password from gaining access to the system.

Now consider DARPA's active authentication program, which seeks to validate users continuously using a wide range of physical and behavioral biometrics such as mouse and typing patterns and how messages and documents are crafted (DARPA 2012). If at any time a user's actions are inconsistent with their normal biometric patterns (called their "cognitive

fingerprint"), access could be terminated. Such a mechanism would be more active than the password mechanism above, as it could keep the adversary from entering and then exploiting any legitimate account on the system. It might even thwart a malware attack, as the malware's behavior would not match that of the account under which it is running.

Consider next a simple firewall access control list (ACL) that blocks all incoming packets to a particular port on the grounds that because the system does not support any services on that port, it would be an open door for attackers. We consider this passive, as it serves more to eliminate a vulnerability than to address a particular threat. However, the ACL would become an element of an active defense if an intrusion prevention system (IPS) detected hostile traffic and then revised the ACL to block the offending traffic. However, an intrusion detection system (IDS) alone is more passive, as it serves primarily as a means of detection and warning.

> active defenses are direct actions taken against specific threats, while passive defenses focus more on protecting cyber assets

Anti-malware (aka anti-virus) tools have much in common with intrusion prevention systems. They detect malicious software, including viruses, worms, and Trojans, and then (optionally) block the code from entering or executing on a protected system. Typically these tools are regularly updated to include signatures for new forms and variants of malware that are detected across the Internet. In this sense, the active defenses are applied globally over the Internet. After new malware is discovered, security vendors create and distribute new signatures to the customers of their anti-malware products.

Intrusion prevention can likewise be performed on a broader scale than a single network or even enterprise. For example, the IP addresses of machines that are spewing hostile packets can be shared widely through "blacklists" and then blocked by Internet service providers. Indeed, victims of massive denial-of-service (DoS) attacks frequently ask upstream service providers to drop packets coming from the originating IP addresses.

Anti-malware and intrusion prevention systems can be integrated to form powerful active defenses. In many respects, the combined defenses would resemble an active air and missile defense system that detects hostile air threats and then takes such actions as shooting them down or jamming their communication, only in cyberspace the defenses are applied to hostile cyber threats such as malicious packets and malware. Rather than targeting incoming ballistic missiles, cyber defenses take their aim at packets that act like "cyber missiles."

Honeypots, which lure or deflect attackers into isolated systems where they can be monitored, are another form of active defense. They are like the decoys used in air defense to deflect missiles away from their intended targets.

In addition to playing a role in network security, active cyber defenses have been used to take down botnets (networks of compromised computers) and counter other cyber threats. The following two examples illustrate.

Coreflood Takedown

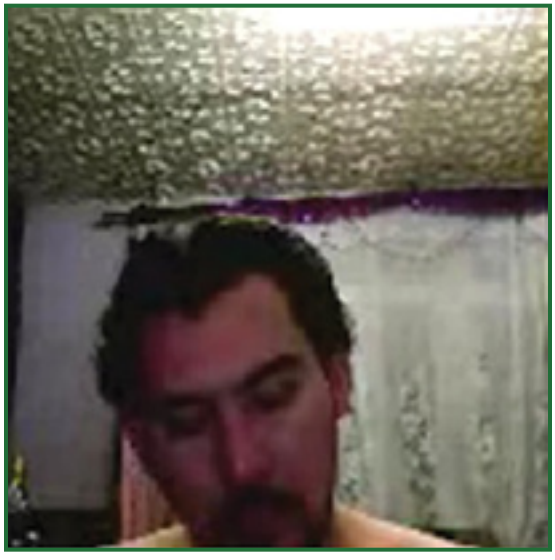
In April 2011, the Federal Bureau of Investigation (FBI), Department of Justice, and the Internet Systems Consortium (ISC) deployed active defenses to take down the Coreflood botnet (Zetter 2011a, 2011b; Higgins 2011). At the time, the botnet comprised over 2 million infected computers, all under the control of a set of command and control (C2) servers. The bot malware installed on the machines was used to harvest usernames and passwords, as well as financial information, in order to steal funds. One C2 server alone held about 190 gigabytes of data stolen from over 400,000 victims.

The active defense included several steps. First, the U.S. District Court of Connecticut issued a temporary restraining order that allowed the non-profit Internet Systems Consortium (ISC) to swap out Coreflood's C2 servers for its own servers. The order also allowed the government to take over domain names used by the botnet. With the infected machines now reaching out to the new C2 servers for instructions, the bots were commanded to "stop." The malware reactivated following a reboot, but each time it contacted a C2 server, it was instructed to stop. The effect was to neutralize, but not eliminate, the malware installed on the compromised machines. To help victims

remove the malware, the FBI provided the IP addresses of infected machines to ISPs so they could notify their customers. In addition, Microsoft issued an update to its Malicious Software Removal Tool, so that victims could get rid of the code.

Using the air defense analogy, the Coreflood takedown can be likened to an active defense against hijacked aircraft, where the hijackers were acting on instructions transmitted from a C2 center. In this situation, the air defense might jam the signals sent from the center and replace them with signals that command the hijackers to land at specified airports. The airports would also be given information to identify the hijacked planes so that when they landed, the hijackers could be removed.

This approach of neutralizing the damaging effects of botnets by commandeering their C2 servers has been used in several other cases. Microsoft, for example, received a court order in November 2012 to continue its control of the C2 servers for two Zeus botnets. Because Zeus had been widely used to raid bank accounts, the operation has no doubt prevented considerable harm (Lemos 2012).



Georgian Outing of Russia-Based Hacker

In October 2012, *Network World* reported that the Georgian government had posted photos of a Russian-based hacker who had waged a persistent, months-long campaign to steal confidential information from Georgian government ministries, parliament, banks, and non-government organization (Kirk 2012). The photos, taken by the hacker's own webcam, came after a lengthy investigation that began in March 2011 when a file on a government computer was flagged by an anti-virus program. After looking into the incident, government officials determined that 300 to 400 computers in key government agencies had been infected with the malware, and that they had acquired it by visiting Georgian news sites that had been infected themselves, in particular, on pages with headlines such as "NATO delegation visit in Georgia" and "U.S.-Georgian agreements and meetings." Once installed, the malware searched for documents using keywords such as USA, Russia, NATO, and CIA, and then transmitted the documents to a drop server where they could be retrieved by the spy.

Georgia's initial response included blocking connections to the drop server and removing the malware from the infected websites and personal computers. However, the spy did not give up and began sending the malware out as a PDF file attachment in a deceptive email allegedly from admin@president.gov.ge.

The Georgian government then let the hacker infect one of their computers on purpose. On that computer, they hid their own spying program in a ZIP archive entitled "Georgian-NATO Agreement." The hacker took the bait, downloaded the archive, and unwittingly launched the government's code. The spyware turned on the hacker's webcam and began sending images back to the government. It also mined the hacker's computers for documents, finding one that contained instructions, in Russian, from the hacker's handler about who to target and how, as well as circumstantial evidence suggesting Russian government involvement.

Again using the air defense analogy, the steps taken to block the exfiltration of files from compromised computers to the drop servers could be likened to jamming the transmission of sensitive data acquired with a stolen reconnaissance plane to the thieves' drop center. The steps taken to bait the hacker into unwittingly stealing and installing spyware might be likened to a command intentionally permitting the theft of a rigged reconnaissance plane with hidden surveillance equipment that sends data it collects about the thieves back to the command.

CHARACTERISTICS AND ETHICAL ISSUES IN ACTIVE CYBER DEFENSE

In this section, we offer a set of distinctions for characterizing the different types of active defense described in the preceding section and discuss some of the ethical issues raised by each.

Scope of Effects

The first set of distinctions pertains to the scope of effects of an active defense. An active defense is said to be internal if the effects are limited to an organization's own internal network. If it affects outside networks, it is said to be external.

Drawing on the air defense analogy, an internal cyber defense is like an air defense system that takes actions against an incoming missile or hostile aircraft after it has entered a country's airspace, while an external cyber defense is like an air defense system that operates in someone else's airspace or attacks the base in a foreign country where the missile is being launched or the hostile aircraft taking off.

We consider defenses that involve sharing threat information with outside parties to be external. An example is the Defense Industrial Base (DIB) Cybersecurity Information Assurance (CS/IA) program operated by the Department of Defense. Under the program, DoD provides DIB companies with unclassified indicators (signatures) of cyber threats. An optional part of the program, called DIB Enhanced Cybersecurity Services (DECS) and run jointly with DHS, allows DoD also to share classified threat information (DoD 2012a).

Most of the effects in the Coreflood takedown were external. In particular, the ISC-operated C2 servers instructed bots in outside networks to stop. In contrast, most of the effects in the Georgian case were internal. Connections to the drop server were blocked on internal networks and internal machines were cleaned of the malware. However, there were also external effects, namely, infection of the hacker's own computer with spyware.

Ethical Issues

In general, most of the ethical issues regarding active defenses concern external active defenses. These will be discussed in the next section when we distinguish cooperative external defenses from non-cooperative ones. However, even internal defenses can raise ethical issues. For example, inside users might complain that their rights to free speech were violated if internal defenses blocked their communications with outside parties. In addition, internal defenses do nothing to mitigate threats across cyberspace. By not even sharing threat information with outsiders, external networks are exposed to continued harm that might be avoided if the defenses were applied to them as well. Arguably, at least in terms of national cyber defense, a better moral choice would be to help mitigate cyber threats more broadly, as DoD has done with its DIB CS/IA and DECS programs. Returning to the air defense analogy, a missile defense system that only shot down missiles headed to military bases would not be as "just" as one that also shot down missiles headed to civilian targets such as cities and malls. On the other hand, it would be unreasonable to expect that missile defense system to protect the air space of other countries, at least absent an agreement to do so.

Degree of Cooperation

The second set of distinctions pertains to the degree of cooperation in an active defense. If all effects against a particular network are performed with the knowledge and consent of the network owner, they are said to be cooperative. Otherwise, they are classified as non-cooperative. For the purpose of discussion here, we assume that network owners are authorized to conduct most defensive operations on their own networks, at least as long as they do not violate any laws or contractual agreements with their customers or users. Thus, the distinction applies mainly to active defenses with external effects.

Using the air defense analogy, a cooperative cyber defense is like an air defense system that shoots down missiles or hostile aircraft in the airspace of an ally that has requested help, while a non-cooperative cyber defense is like an air defense system that shoots them down in the adversary's own airspace.

Anti-viral tools are cooperative defenses. Security vendors distribute new signatures to their customers, but the signatures are only installed with customer permission. Similarly, sharing blacklists of hostile IP addresses is cooperative. In general, any active defense that does nothing more than share threat information is cooperative.

Defenses become non-cooperative when they involve actions taken against external computers without permission of the user or network owner. In the case of Coreflood, the actions taken against the individual bots were non-cooperative. Neither the users of those machines nor the owners of the networks on which they resided agreed to have the bot code stopped. On the other hand, neither had they agreed to the initial malware infection and subsequent theft of their data. Arguably, any user would prefer that the malware be stopped rather than be allowed to continue its harmful actions. Further, even though the action was non-cooperative, it was deployed under legal authorities, enabled in part by the temporary restraining order. Moreover, the actual elimination of the malware from the infected machines was to be a cooperative action involving the machine owners.

Non-cooperative defenses include what is sometimes called "attack back," "hack back," or "counterstrike" where the defense uses hacking or exploit tools directly against the source of an attack or gets the attacker to unwittingly install software, say by planting it in a decoy file on a computer the attacker has compromised. The goal might be to collect information about the source of the attack, block attack packets, or neutralize the source. Non-cooperative defenses also include court-ordered seizures of computers.

Although the Coreflood takedown did not include any sort of hack back, the Georgian case did. In particular, the actions taken to plant spyware on the hacker's computer constituted a non-cooperative counterstrike. However, one could argue that the hacker would never have acquired the spyware had he not knowingly and willfully first infected the computer hosting it and second downloaded the ZIP archive containing it. Thus, he was at least complicit in his own infection and ultimate outing.

Ethical Issues

As a rule, non-cooperative defenses, particularly those involving some sort of hack back, raise more ethical and legal issues than cooperative ones. In part, this is because most cyber attacks are launched through machines that themselves have been attacked, making it hard to know whether the immediate source of an attack is itself a victim rather than the actual source of malice. They may be hacked servers or bots on a botnet. Thus, any actions taken against the computers could harm parties who are not directly responsible for the attacks. In addition, cyber attacks in general violate computer crime statutes, at least when conducted by private sector entities. While the argument can be made that some hack backs would be permissible under the law, not everyone agrees, and the topic has been hotly debated (Denning 2008, Steptoe 2012, Messmer 2012). However, government entities, in particular the military and law enforcement and intelligence agencies, have or can acquire the authorities needed to perform actions that might be characterized as hacking under certain prescribed conditions.

If we assume that non-cooperative defenses are conducted by or jointly with government entities with the necessary legal authorities, then the primary concern is that innocent parties may be harmed. Then we can draw on the long tradition of just war theory to determine the conditions under which active cyber defenses that pose risks to noncombatants can be ethically justified.

Most just war theorists hold that noncombatant immunity is a key linchpin to all our moral thinking in war (Walzer 1977, Nagel 1972, Rodin 2003, Orend 2006). As such, noncombatants are never to be intentionally targeted for harm as any part of a justified military action. Traditional just war theory does hold, however, that some actions that will foreseeably but unintentionally harm noncombatants may be permissible, so

long as that harm is truly unintentional, is proportionate to the good goal achieved by the act, and is not the means itself to achieve the good goal. Grouped together, these principles are known as the doctrine of double effect. The doctrine has come under heavy scholarly debate, with many critics doubting that its principles can hold true for all cases (Davis 1984, Kamm 2004, McIntyre 2001, Steinhoff 2007), while others have argued that some revised or narrowed version of the doctrine can still be defended and applied to war (McMahan 1994, Quinn 1989, Nelkin and Rickless 2012). We cannot here engage this larger debate, but assume that at least some narrow version of the doctrine of double effect is true and, as such, it is critical for our moral conclusions regarding harm to noncombatants from active cyber defense.

Whether noncombatants' property can be targeted is another matter. Generally, noncombatant property is similarly considered immune from direct and intentional harm since harming one's property harms that person. However, as with physical harm, unintended harm of noncombatant property can be permissible in some instances. Moreover, traditional just war theory and the laws of armed conflict can allow for some level of intentional harm to civilian property if it is necessary to block a particularly severe enemy military action and the civilians in question are later compensated. That is, generally, the ethical restrictions on harm to civilian property are far less strict than for physical harm to civilian persons. This is true for unintentional harms of both kinds, and can even allow for some intentional harm to property when necessary, the stakes are high enough, and recompense can be made.

In the case of active air defense, systems like Iron Dome are not without risk to civilians. If someone happens to be under an incoming rocket's flight path at the time it is hit, they could be harmed from fallout from the explosion. However, Israel has limited their counterstrikes primarily to rockets aimed at densely populated urban areas. In that situation, any fallout is likely to be substantially less harmful than the effects produced by the rockets themselves if allowed to strike. We argue that such a risk imposition can be morally warranted. Note, however, that if Iron Dome created large amounts of dangerous and lethal fallout disproportionate to the lives saved, then its use would not be permissible.

In general, if an air defense system distributes some small level of risk of harm to civilians under an incoming missile's flight path in order to protect a much larger number of civilians from much greater harm, then the conditions are present for such defense to be morally permissible. This is precisely what we find in the case of real-world air defense systems such as Iron Dome. Further, it is irrelevant whether the risk of harm is imposed on noncombatants from one's own state or another state. The reason is that what matters are the moral rights of *all* noncombatants, including, of course, noncombatants on either side of a given conflict. The point is to minimize collateral harm to all noncombatants.

The same principles should apply to active cyber defense; that is, it should be morally permissible for a state to take an action against a cyber threat if the unjust harm prevented exceeds and is proportionate to any foreseen harm imposed on noncombatants. Indeed, in the cyber domain it will often be easy to meet this demand because it is often possible to shoot down the cyber missiles without causing any fallout whatsoever. Instead, packets are simply deleted or diverted to a log file. Nobody is harmed.

In some cases, however, an active defense could have a negative impact on innocent parties. To illustrate, suppose that an action to shut down the source of an attack has the effect of shutting down an innocent person's computer that had been compromised and used to facilitate the attack. In this case, the action might still be morally permissible. There are two reasons. First, the harm induced might be temporary in

nature, affecting the computer, but only for a short time until the attack is contained. Second, the harm itself might be relatively minor, affecting only noncombatants' property, not their persons. It is possible that such effects could further impede other rights of the noncombatants, such as their ability to communicate or engage in activity vital to their livelihoods. But all of these further harms would be temporary in nature and could even be compensated for, if appropriate, after the fact. This is not to disregard the rights of noncombatants and their property and its use for the furtherance of other rights in our moral calculus, but is rather a simple recognition that different kinds and severities of harm result in different moral permissions and restrictions.

The fact that the harm itself is likely to be non-physical is quite significant in our moral reasoning in favor of active cyber defense. If it is permissible in some cases to impose the risk of *physical* harm on noncombatants as part of a necessary and proportionate defensive action against an incoming missile (as we argued above that it could be in the air defense case), then surely there will be cases where it can be permissible to impose the risk of temporary harm to the *property* of noncombatants in order to defend against an unjust cyber attack. The point here with active cyber defense is that the *kind* of harms that would be potentially imposed on noncombatants, in general, are the kinds of reduced harms that should make such defensive actions permissible.

A caveat, however, is in order. Computers today are used for life-critical functions, for example, to control life support systems in hospitals and operate critical infrastructure such as power grids. In a worst-case, an active cyber defense that affects such a system might lead to death or significant suffering. These risks need to be taken into account when weighing the ethics of any non-cooperative action that could affect noncombatants. In general, defensive actions that do not disrupt legitimate functions are morally preferable over those that do. If the scope of possible effects cannot be reasonable estimated or foreseen, then the action may not be permissible.

In the case of Coreflood, the takedown affected many noncombatant computers. However, the effect was simply to stop the bot code from running. No other functions were affected, and the infected computer continued to operate normally. Thus, there was virtually no risk of causing any harm whatsoever, let alone serious harm. In the Georgian case, the only harm was to the attacker's own computer—and he brought this on himself by downloading the bait files, thus making himself liable to intentional defensive harm.

Although the discussion here has focused on non-cooperative defenses, it is worth noting that while cooperative defenses generally raise fewer issues, they are not beyond reproach. For example, suppose that a consortium of network owners agrees to block traffic from an IP address that is the source of legitimate traffic as well as the hostile traffic they wish to stop. Depending on circumstances, a better moral choice might be to block only the hostile traffic or work with the owner of the offending IP address to take remedial action.

While cooperative defenses generally raise fewer issues, they are not beyond reproach.

Types of Effects

The third set of distinctions pertains to the effects produced. An active defense is called sharing if the effects are to distribute threat information such as hostile IP addresses or domain names, or signatures for malicious packets or software, to other parties. Sharing took place in the Coreflood takedown when the FBI provided the IP addresses of compromised machines within the United States to their U.S. ISPs and to foreign law enforcement agencies when the machines were located outside the U.S.. Another example of sharing is DoD's DIB program, described earlier.

An active defense is called collecting if it takes actions to acquire more information about the threat, for example, by activating or deploying additional sensors or by serving a court order or subpoena against the source or an ISP likely to have relevant information. In the Coreflood takedown, the replaced C2 servers were set up to collect the IP addresses of the bots so that eventually their owners could be notified. The servers did not, however, acquire the contents of victim computers. In the Georgian case, spyware was used to activate a webcam and collect information from the attacker's computer.

An active defense is called blocking if the effects are to deny activity deemed hostile, for example, traffic from a particular IP address or execution of a particular program. The Coreflood takedown had the effect of breaking the communication channel from the persons who had been operating the botnet to the C2 servers controlling it. As a result, they could no longer send commands to the bots or download stolen data from the servers. In the Georgian case, connections to the drop servers were blocked in order to prevent further exfiltration of sensitive data.

Finally, an active defense is pre-emptive if the effects are to neutralize or eliminate a source used in the attacks, for example, by seizing the computer of a person initiating attacks or by taking down the command and control servers for a botnet. In the Coreflood takedown, the hostile C2 servers were put out of commission and the bots neutralized. With further action on the part of victims, the malware could also be removed.

Using the air defense analogy, the cyber defense of sharing is like a missile defense system that reports new missile threats to allies so that they can shoot them down. The cyber defense of collecting is like a missile defense system that activates additional radars or other sensors in response to an increased threat level, or that sends out sorties to investigate suspicious aircraft. The cyber defense of blocking is akin to a missile defense system that shoots down incoming missiles or jams their radars and seekers. Finally, the cyber defense of pre-emption is like launching an offensive strike against the air or ground platform launching the missiles.

Some authors regard retaliation or retribution as a form of active defense. However, we consider these operations to be offensive in nature, as they serve primarily to harm the source of a past attack rather than mitigate, stop, or pre-empt a current one.

Ethical Issues

All four types of cyber operations raise ethical issues. The act of sharing raises issues of privacy and security, particularly if any sensitive information is shared. The act of collecting also raises issues about privacy and security, but in this case relating to the new information acquired rather than the dissemination of existing information. The act of blocking raises issues relating to free speech and over-blocking. In a worst case, traffic might be blocked that is important for the operation of a life-support system or critical infrastructure such as power generation and distribution. Likewise, the act of pre-emption raises ethical issues relating to disabling software or systems. Again, a worst-case scenario could cause serious harm, for example, by shutting down a life-support system. These possible harms would need to be considered in the application of any non-cooperative cyber defense, as discussed in the previous section, and argue for defenses that limit their effects, say, by disabling only traffic and software involved in an attack rather than shutting down all traffic and complete systems.

In the Coreflood takedown, it is important to note that the government did not attempt to remove the bot code from infected machines. They only neutralized it by issuing the stop command. Part of the reason for not removing the code was a concern for unanticipated side effects that might damage an infected computer.

Because active cyber defense is a form of defense that should not be misconstrued as offense, it is worth explaining why the distinction between offensive retaliation versus legitimate defensive action is so crucial in the ethical dimensions of killing and war. Defensive harm has the lowest ethical barrier to overcome from amongst all possible justifiable harms. That is, if one is being wrongly attacked, then the moral restrictions against using force of some kind in order to block that wrongful attack are (relatively) few. This is because all people have a

right not to be harmed unjustly. If one is attempting to harm someone unjustly, then she has made herself morally liable to suffer defensive harm as part of an act taken to thwart her attempted unjust harm. The person being wrongly attacked may permissibly harm his attacker in an effort to block or thwart the attack against him, so long as the defensive harm meets two criteria. First, it must be necessary to inflict the defensive harm to block the unjust attack. If the defensive harm in question does nothing to block the liable party's unjust attack, then it is retributive punishment, or something else, but not properly an act of defense. Second, the defensive harm must be proportionate to the unjust harm to be blocked. If a foreign plane was found conducting reconnaissance over a state's territory without permission during peacetime, then the foreign state may have made itself liable to some form of defensive action such as being escorted to an airfield. However, it would be disproportionate and wrongful to shoot the plane down or, even worse, shoot down commercial planes flying under the foreign state's flag. In general, there must be some reasonable correlation and proper "fit" between the extent of defensive response and the degree of liability of the party defended against (McMahan 2005, Quong 2012). In the case of an active cyber defense, if the act is truly a defensive effort to block an unjust attack, then so long as it is necessary and proportionate, it will usually be ethically permissible. In the Georgian case, the government responded to the cyber espionage operation against it with its own espionage operation against the hacker. It did not destroy software and data on the hacker's computer.

Degree of Automation

The final set of distinctions pertains to the degree of human involvement. An active defense is said to be automatic if no human intervention is required and manual if key steps require the affirmative action of humans.

Most anti-malware and intrusion prevention systems have both manual and automated components. Humans determine what goes into the signature database, and they install and configure the security software. However, the processes of signature distribution, malicious code and packet detection, and initial response are automated.

In the Coreflood takedown, the execution of the stop commands was fully automated through the C2 servers. However, humans played an important role in planning and decision making, analyzing the botnet code and the effects of issuing a stop command, acquisition of the restraining order, and swapping out of the C2 servers. Thus, the entire operation had both manual and automatic aspects. In the Georgian case, much of the investigation involved manual work, including analyzing the code, determining what the hacker was looking for, and setting up the bait with the spyware. But the key element in the outing, namely the operation of the spyware, was automated. Once the hacker downloaded the ZIP archive, it did the rest.

Applying the air defense analogy once again, an automatic cyber defense is like a missile defense system that automatically shoots down anything meeting the preset criteria for being a hostile aircraft or incoming missile, whereas a manual cyber defense is more like Operation Noble Eagle where humans play a critical role, both in recognizing and responding to suspicious activity in U.S. airspace.

Ethical Issues

In general, manual actions give humans a greater opportunity to contextualize their ethical decisions. Rather than configuring a system to respond always in a certain way, humans can take into account the source or likely source of a perceived threat, its nature, and the likely consequences of taking certain actions against it. This is vital to Nobel Eagle, where most incidents

turn out to be non-hostile and lives are at stake. On the other hand, manual actions take longer to execute than automated ones, potentially allowing greater damage to be incurred before the threat is mitigated. In the cyber domain, where actions can take place in an instant, automated defenses become critical. That is, the speed of some actions in the cyber domain are such that a cyber defense must be automated in order to have any effect at all against the attack. It is perhaps for this reason that some cyber actions have been given an exemption from the recent "man in the loop" legal requirements for automated weapon systems put out by the DoD (DoD 2012b, Gallagher 2012). If a hostile actor has launched an attack to cause a power generator to explode, then an automated response that successfully blocks the attack without causing unnecessary harm is morally superior to a manual one that comes too late.

However, this does not mean that all cyber defenses should be automated. To be clear: we are not arguing that *all* cyber actions should be exempt from the "man in the loop" requirement. The nature of a defense and its potential effects must be weighed in any decision to automate. The potential severity of foreseeable harms should govern whether it should be automated. It is true that the cyber case is unique in that the speed of many cyber attacks necessitates that many defenses be automated in order to be effective in any way. But if the effects of a given defense are such that their automation would lead to too great a risk of impermissible harm, then they should not be automated, even if this entirely nullifies their efficacy. Thankfully, given the reasons discussed above regarding the kinds of predictable effects that most forms of active cyber defense would result in, we find that in many cases their automation could be permissible.

Conclusions

Using analogies from air defense, we have shown that active cyber defense is a rich concept that, when properly understood and executed, is neither offensive nor necessarily harmful and dangerous. Rather, it can be executed in accordance with the well-established ethical principles that govern all forms of defense, namely principles relating to harm, necessity, and proportionality. In many cases, such as with most botnet takedowns, active defenses mitigate substantial harm while imposing little or none of their own.

While active defenses can be morally justified in many cases, we do not mean to imply that they always are. All plausible effects must be considered to determine what, if any, harms can follow. If harms cannot be estimated or are unnecessary or disproportionate to benefits gained, an active defense cannot be morally justified.

In considering active defenses, we have assumed that they would be executed under appropriate legal authorities. In particular, they would be conducted by authorized government entities or by private companies operating under judicial orders or otherwise within the law. We leave open the question of how far companies can go in areas where the law is unclear or untested. While such active defenses as sharing attack signatures and hostile IP addresses and domain names have raised few legal questions, an active defense that deleted code or data on the attacker's machine would raise more. No doubt, this area will likely continue to inspire lively discussion and debate. ❊

REFERENCES

Ackerman, G. and Ramadan, S. A. (2012) 'Israel Wages Cyber War With Hamas as Civilians Take Up Computers,' *Bloomberg*, November 19. http://www.bloomberg.com/news/2012-11-19/israel-wages-cyber-war-with-hamas-as-civilians-take-up-computers.html (accessed November 26, 2012).

Air Force. (2012) Operation Noble Eagle, Air Force Historical Studies Office, Posted September 6. http://www.afhso.af.mil/topics/factsheets/factsheet.asp?id=18593 (accessed November 6, 2012).

DARPA. (2012) 'Active Authentication,' DARPA Information Innovation Office, http://www.darpa.mil/Our_Work/I2O/Programs/Active_Authentication.aspx (accessed November 6, 2012).

Davis, N. (1984) 'The Doctrine of Double Effect: Problems of Interpretation,' *Pacific Philosophical Quarterly* 65: 107–123.

Denning, D. E. (2008) 'The Ethics of Cyber Conflict,' Chapter 17 in *The Handbook of Information and Computer Ethics* (K. E. Himma and H. T. Tavani eds.), Wiley, pp. 407–428.

Dittrich, D. and Himma, K. E. (2005) 'Active Response to Computer Intrusions,' *The Handbook of Information Security* (Bidgoli, H. ed.), John Wiley & Sons.

DoD. (2012a) Fact Sheet: Defense Industrial Base (DIB) Cybersecurity Activities, May 11, 2012. http://www.defense.gov/news/d20120511dib.pdf (accessed December 5, 2012).

DoD. (2012b) Directive Number 3000.09, 'Autonomy in Weapon Systems,' November 21, 2012. www.dtic.mil/whs/directives/corres/pdf/300009p.pdf (accessed December 6, 2012).

Gallagher, S. (2012) 'U.S. cyber-weapons exempt from "human judgment" requirement' *arstechnica,* November 29, 2012. http://arstechnica.com/tech-policy/2012/11/us-cyber-weapons-exempt-from-human-judgment-requirement/ (accessed December 5, 2012).

Higgins, K. J. (2011) 'Coreflood Botnet An Attractive Target For Takedown For Many Reasons,' *Dark Reading*, April 14. http://www.darkreading.com/database-security/167901020/security/client-security/229401635/coreflood-botnet-an-attractive-target-for-takedown-for-many-reasons.html (accessed November 27, 2011).

JP 3-01 (2012) 'Countering Air and Missile Threats,' Joint Publication 3-01, March 23.

Kamm, F. (2004) 'Failures of Just War Theory: Terror, Harm, and Justice,' *Ethics* 114: 650–92.

Kershner, I. (2012) 'Israeli Iron Dome Stops a Rocket With a Rocket,' *The New York Times*, November 18. http://www.nytimes.com/2012/11/19/world/middleeast/israeli-iron-dome-stops-a-rocket-with-a-rocket.html?_r=0 (accessed November 19, 2012).

Kirk, J. (2012) 'Irked By Cyberspying, Georgia Outs Russia-Based Hacker—With Photos,' *Network World*, October 30. http://www.networkworld.com/news/2012/103012-irked-by-cyberspying-georgia-outs-263790.html (accessed November 27, 2012).

Lemos, R. (2012) 'Microsoft Can Retain Control of Zeus Botnet Under Federal Court Order,' eWeek, December 1. http://www.eweek.com/security/microsoft-can-retain-control-of-zeus-botnet-under-federal-court-order/ (accessed December 7, 2012).

McIntyre, A. 'Doing Away with Double Effect,' *Ethics* 111: 219–55.

McMahan, J. (1994) 'Revising the Doctrine of Double Effect,' The Journal of Applied Philosophy 11(2): 1993-221.

McMahan, J. (2005) 'The Basis of Moral Liability to Defensive Harm,' *Philosophical Issues* 15: 386–405.

Messmer, E. (2012) 'Hitting Back at Cyberattackers: Experts Discuss Pros and Cons,' *Network World*, November 1. http://www.networkworld.com/news/2012/110112-cyberattackers-263885.html (accessed November 29, 2012).

Nagel, T. (1972) 'War and Massacre,' *Philosophy and Public Affairs*, 1(2): 123–144.

Nelkin, D. and Rickless, S. (2012) 'Three Cheers for Double Effect,' *Philosophy and Phenomenological Research,* http://onlinelibrary.wiley.com/doi/10.1111/phpr.12002/full (accessed December 5, 2012).

Orend, B. (2006) *The Morality of War*, Peterborough, ON: Broadview Press.

Patriot. 'MIM-104 Patriot,' Wikipedia. http://en.wikipedia.org/wiki/MIM-104_Patriot (accessed November 6, 2012).

Quinn, W. S. (1989) 'Actions, Intentions, and Consequences: The Doctrine of Double Effect,' *Philosophy and Public Affairs* 18:334–51.

Quong, J. (2012) 'Liability to Defensive Harm,' *Philosophy & Public Affairs*, 40(1): 45–77.

Rodin, D. (2003) *War and Self-Defence*, New York: Oxford University Press.

Steinhoff, U. (2007) *On the Ethics of War and Terrorism*, Oxford: Oxford University Press.

Steptoe. (2012) 'The Hackback Debate,' Steptoe Cyberblog, November 2. http://www.steptoecyberblog.com/2012/11/02/the-hackback-debate/ (accessed November 29).

Walzer, M. (1977) *Just and Unjust Wars*. New York: Basic Books

Zetter, K. (2011a) 'With Court Order, FBI Hijacks Coreflood Botnet, Sends Kill Signal," *Wired*, April 13. http://www.wired.com/threatlevel/2011/04/coreflood/ (accessed November 27, 2012).

Zetter, K. (2011b) 'FBI vs. Coreflood Botnet: Round 1 Goes to the Feds,' *Wired*, April 26. http://www.wired.com/threatlevel/2011/04/coreflood_results/ (accessed November 27, 2012).

NOTES

1 Approved for public release; distribution is unlimited. The views expressed in this document are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.