Introduction

At least on the surface, most cyber attacks appear to be clearly unethical as well as illegal. These include attacks performed for amusement or bragging rights, such as web defacements conducted “just for fun” and computer viruses launched out of curiosity but disregard for their consequences. They also include attacks done for personal gain, such as system intrusions to steal credit card numbers and trade secrets; denial-of-service attacks aimed at taking out competitor websites or extorting money from victims; and attacks that compromise and deploy large “botnets” of victim computers to send out spam or amplify denial-of-service attacks.

There are, however, three areas of cyber conflict where the ethical issues are more problematic. The first is cyber warfare at the state level when conducted in the interests of national security. Some of the questions raised in this context include: Is it ethical for a state to penetrate or disable the computer systems of an adversary state that has threatened its territorial or political integrity? If so, what are the ground rules for such attacks? Can cyber soldiers attack critical infrastructures such as telecommunications and electric power that serve both civilian and military functions? If a nation is under cyber assault from another country, under what conditions can it respond in kind or use armed force against the assailant? Can it attack computers in a third country whose computer networks have been compromised or exploited to facilitate the assault?

The second area with ethical dilemmas involves non-state actors whose cyber attacks are politically or socially motivated. This domain of conflict is often referred to as “hacktivism,” as it represents a confluence of hacking with activism. If the attacks are designed to be sufficiently destructive as to severely harm and terrorize civilians, they become “cyberterrorism” – the integration of cyber attacks with terrorism. While cyberterrorism is abhorrent and clearly unethical, hacktivism raises ethical questions. For example: Is it ethical for a group of hackers to take down a website that is being used primarily to trade child pornography, traffic in stolen credit card numbers, or support terrorist operations? Can the hacktivists protest the policies or practices of governments or corporations by defacing websites or conducting web “sit-ins?” Can they attack vulnerable machines in order to expose security holes with the goal of making the Internet more secure?

Finally, the third area involves the ethics of cyber defense, particularly what is called “hack back,” “strike back,” or “active response.” If a system is under cyber attack, can the system administrators attack back in order to stop it? What if the attack is
coming from computers that may themselves be victims of compromise? Since many attacks are routed through chains of compromised machines, can a victim “hack back” along the chain in order to determine the source?

This paper explores ethical issues in each of these areas of cyber conflict. The objective is not to answer the questions listed above, but rather to offer an ethical framework in which they can be addressed. Examples are used to illustrate the principles, but no attempt is made to reach a final ethical decision. To do so would require a much more thorough analysis of the nature of a particular cyber attack and the context in which it is used.

The framework presented here is based on the international law of armed conflict. Although this law was developed to address armed attacks and the use of primarily armed force, some work has been done to interpret the law in the domain of cyber conflict. The law has two parts: *jus ad bellum*, or the law of conflict management, and *jus in bello*, or the law of war. Despite being referred to as “law,” both of these are as much about ethical behavior as they are rules of law.

The international law of armed conflict applies to nation states, and thus concerns cyber warfare at the state level. The paper will extend this framework to politically and socially motivated cyber attacks by non-state actors, and compare this approach with some previous work on the ethics of cyber activism and civil disobedience. It will also apply the international law of armed conflict to the domain of cyber defense, and show how it ties in with the legal doctrine of self-defense and relates to other work on hack back.

Thus, for all three areas, the paper builds on the ethical principles encoded in the international law of armed conflict, and interpretation of those principles in the cyber domain. In this way, the paper approaches the three areas of cyber attack more as domains of conflict, especially international conflict, than as domains of crime—even though the acts themselves may also violate criminal statutes.

There are several areas of cyber conflict the paper does not address. Besides cyber attacks conducted for pleasure or personal gain, the paper does not consider revenge attacks by insiders—all of which are generally regarded as unethical. In addition, the paper does not address methods of cyber conflict other than cyber attacks, for example, messages transmitted for the purpose of psychological operations or deception. While these other types of activity raise important ethical issues, their treatment is beyond the scope of this paper.

**Cyber Warfare at the State Level**

The law of international conflict consists of two parts: *jus ad bellum*, or the law of conflict management, and *jus in bello*, or the law of war. Both are concerned with the
use of force, particularly armed forces, but the former specifies \textit{when} that force may
be applied, while the latter specifies ground rules for \textit{how} it should be applied. Both
are about ethical principles as much as they are about “law,” and indeed, international
law does not carry the same weight as domestic law. Under international law, states,
as sovereign entities, assume international legal obligations only by affirmatively
agreeing to them, for example, signing a treaty or agreeing to abide by the Charter of
the United Nations. They are free to decline participation, and they are free to back
out later. By contrast, under domestic laws, the citizens of a country are vulnerable to
prosecution for violating any laws, regardless of whether they agree with them, and
regardless of whether the laws are even just.

The law of international conflict is designed to promote peace and minimize the
adverse effects of war on the world. As a general rule, states are not permitted to
attack other states, except as a means of self-defense. Where conflict does arise, the
law is intended to ensure that wars are fought as humanely as possible, minimizing
collateral damage (harm to civilians and civilian property). Thus, the international
law of armed conflict tends to prescribe widely accepted ethical principles.

\textit{Jus ad Bellum – The Law of Conflict Management}

The law of conflict management is primarily concerned with the application of force,
particularly armed force. It is codified in the United Nations Charter and specifies the
conditions under which member states may apply force against other states. The most
relevant parts of the Charter are Articles 2(4), 39, and 51.

Article 2(4) prohibits states from using force against other states:

\begin{quote}
All Members shall refrain in their international relations from the threat or use
of force against the territorial integrity or political independence of any state,
or in any other manner inconsistent with the Purposes of the United Nations.
\end{quote}

Although the nature of this force is left somewhat open, it may include more than just
the use of armed force, as other parts of the Charter explicitly refer to armed force.
However, it is not so broad as to cover generally lawful activity such as boycotts,
economic sanctions, severance of diplomatic relations, and interruption of
communications. (Wingfield, 2000, p. 90)

Article 39 assigns the U.N. Security Council responsibility for responding to threats
and acts of aggression:

\begin{quote}
The Security Council shall determine the existence of any threat to the peace,
breach of the peace, or act of aggression and shall make recommendations, or
decide what measures shall be taken in accordance with Articles 41 and 42, to
maintain or restore international peace and security.
\end{quote}
Although a wide variety of acts can plausibly be interpreted as a “threat to the peace,” the term “aggression” is defined in a U.N. General Assembly resolution as “the use of armed force” by a member or non-member state. It includes invasions, attacks, bombardments, and blockades by armed military forces and other groups, including mercenaries. Article 41 refers to responses other than armed force, for example “complete or partial interruption of economic relations and means of communication, and the severance of diplomatic relations.” Article 42 refers to the use of air, sea, and land forces, including demonstrations, blockades, and other operations.

While Article 2(4) prohibits states from launching offensive attacks, Article 51 acknowledges a right to self-defense against armed attacks:

Nothing in the present Charter shall impair the inherent right of individual or collective self-defence if an armed attack occurs against a Member of the United Nations, until the Security Council has taken measures necessary to maintain international peace and security. Measures taken by Members in the exercise of this right of self-defence shall be immediately reported to the Security Council and shall not in any way affect the authority and responsibility of the Security Council under the present Charter to take at any time such action as it deems necessary in order to maintain or restore international peace and security.

Although Article 51 states that defensive measures, including the use of force, are allowed after a state has been attacked, it is generally understood that states also have a right of “anticipatory self-defense;” that is, they can take pre-emptive action to avert a strike. They are also permitted to exercise “self-defense in neutral territory.” This means they can use force against a threat operating in a neutral state when that state is unwilling or unable to prevent the use of its territory as a base or sanctuary for attacks. (DoD OGC, 1999, p. 14)

In summary, the UN Charter prohibits states from using force against other states (Article 2(4)), except when conducted in self defense (Article 51) or under the auspices of the Security Council (Article 39). The Charter effectively encodes an ethical principle of just cause for attacking another state that most people would accept. States have a moral right to defend themselves against acts and threats of aggression, but they do not have the right to engage in unprovoked aggression. The use of force is permissible only as a means of defending against aggression.

In order to apply these legal/ethical principles to cyber warfare, we must first determine whether cyber attacks constitute the use of force. If they do, then they would fall under the UN Charter along with armed force, implying that cyber attacks at the state level would be justified only as a means of defense. But if they are not considered to be a form of force, the ethical issues regarding their application are more ambiguous, falling closer to the issues raised by “softer” forms of coercion such as trade restrictions and severance of diplomatic relations.
When Does a Cyber Attack Constitute the Use of Force?

Not all cyber attacks are equal. The impact of a cyber attack that denies access to a news website for one hour would be relatively minor compared to one that interferes with air traffic control and causes planes to crash. Indeed, the effects of the latter would be comparable to the application of force to shoot down planes. Thus, what is needed is not a single answer to the question of whether cyber attacks involve the use of force, but a framework for evaluating a particular attack or class of attacks.

For this, we turn to the work of Michael Schmitt, Professor of International Law and Director of the Program in Advanced Security Studies at the George G. Marshall European Center for Security Studies in Germany. In a 1999 paper, Schmitt, formerly a law professor at both the US Naval War College and the US Air Force Academy, offered seven criteria for distinguishing operations that use force from economic, diplomatic, and other soft measures. (Schmitt, 1999) For each criterion, there is a spectrum of consequences, the high end resembling the use of force and the low end resembling soft measures. The following description is based on both Schmitt’s paper and the work of Thomas Wingfield, author of *The Law of Information Conflict*. (Wingfield, 2000, 120-127)

1. **Severity.** This refers to people killed or wounded and property damage. The premise is that armed attacks that use force often produce extensive casualties or property damage, whereas soft measures do not.

2. **Immediacy.** This is the time it takes for the consequences of an operation to take effect. As a general rule, armed attacks that use force have immediate effects, on the order of seconds to minutes, while softer measures, such as trade restrictions, may not be felt for weeks or months.

3. **Directness.** This is the relationship between an operation and its effects. For an armed attack, effects are generally caused by and attributable to the application of force, whereas for softer measures there could be multiple explanations.

4. **Invasiveness.** This refers to whether an operation involved crossing borders into the target country. In general, an armed attack crosses borders physically, whereas softer measures are implemented from within the borders of a sponsoring country.

5. **Measurability.** This is the ability to measure the effects of an operation. The premise is that the effects of armed attacks are more readily quantified (number of casualties, dollar value of property damage) than softer measures, for example severing diplomatic relations.
6. **Presumptive Legitimacy.** This refers to whether an operation is considered legitimate within the international community. Whereas the use of armed force is generally unlawful absent some justifiable reason such as self-defense, the use of soft measures are generally lawful absent some prohibition.

7. **Responsibility.** This refers to the degree to which the consequence of an action can be attributed to a state as opposed to other actors. The premise is that armed coercion is within the exclusive province of states and is more susceptible to being charged to states, whereas non-state actors are capable of engaging in such soft activity as propaganda and boycotts.

To see how these criteria could apply to a cyber attack, consider an intrusion into an air traffic control system that causes two large planes to enter the same airspace and collide, leading to the deaths of 500 persons on board the two aircraft. In terms of severity, the cyber attack clearly ranks high. Immediacy is also high, although the delay between the intrusion and the crash may be somewhat longer than between something like a missile strike and the planes crashing. With respect to directness, let us assume the reason for the crash is clear from information in the air traffic control computers and the black boxes on board the planes, so directness ranks high. Invasiveness, however, is moderate, requiring only an electronic invasion rather than a physical one. Measurability, on the other hand, is high: 500 people dead and two planes destroyed. Presumptive legitimacy is also high in that the act would be regarded as illegitimate, akin to a missile attack (the high end of the spectrum corresponds to high illegitimacy). Responsibility comes out moderate to high. In principle, the perpetrator could be anyone, but the level of skill and knowledge required to carry out this attack would rule out most hackers, suggesting state sponsorship. In summary, five criteria (severity, immediacy, directness, presumptive legitimacy, and measurability) rank high, while two rank at least moderate (invasiveness and responsibility). Thus, the attack looks more like the application of force than a softer, more legitimate form of coercion.

Now, consider a massive distributed denial of service (DDoS) attack against a key government website that exploits a botnet of hundreds or thousands of compromised computers (“zombies”) and makes the site inaccessible for one day. This would likely rank low to moderate on severity, but high on immediacy. Directness would be moderate to high. Although the effects could as easily be attributed to hardware or software malfunction, network monitoring and inspection of Internet logs would show the problem to be caused by a massive onslaught of traffic. Invasiveness would be about the same as in the previous scenario, namely moderate owing to the electronic penetration. Measurability would be high, as it is easy to determine the downtime of the target web server. Presumptive legitimacy would also score high, as DoS attacks, like force, are generally regarded as illegitimate and in violation of laws. Responsibility would be low to moderate. Some skill is required, but attribution would be difficult and many hackers would be capable of pulling it off. In summary,
the attack looks less like force than the one causing the plane crash in terms of severity and responsibility, but neither does it resemble legitimate measures.

Wingfield suggests assigning a score for each criterion, say from 0 to 10. The idea is that high scores resemble force, whereas low scores resemble the softer measures such as economic and political ones. Under a “primary Schmitt analysis,” the seven scores are summed and the average taken. For a “secondary Schmitt analysis,” the criteria are assigned weights and the weighted average computed. This would allow severity, for example, to count more than the other criteria. An example with graphs showing the results of both primary and secondary Schmitt analyses is given in (Michael, 2003).

To the extent that a particular cyber attack looks like the application of force, its application would violate Article 2(4), possibly triggering an Article 39 response from the UN Security Council or an Article 51 application of force in self-defense by the target. However, under Articles 39 and 51, cyber attacks that resemble force would be allowed as a means of defense against aggressors who use either physical or cyber force.

On the other hand, if the attack looks more like legitimate, soft measures, than the use of force, then its application should not constitute a violation of Article 2(4). Moreover, if not deemed serious, it would likely not trigger an Article 39 response by the UN Security Council, as it would not be interpreted as a threat to the peace or act of aggression. Nor would it provide grounds for the target country to use force in its self-defense under Article 51. Of course, all this is theory. In practice, a nation that is the victim of a cyber attack may perceive it as an act of force worthy of a physical (or cyber) response, regardless of how the perpetrators score it under Schmitt’s criteria or any others.

The ethical implications are that cyber attacks that resemble force are, like the use of physical force, morally justified only when they adhere to Articles 2(4), 39, and 51 of the UN Charter; that is, they are inherently defensive in nature. Unprovoked acts of aggression in cyberspace that resemble the use of force are not legally permissible.

Cyber attacks that fall below the Article 2(4) threshold for force are more likely to be ethical than attacks that cross the threshold, but they are not necessarily morally right. Their ethical implications must be examined like any other government action, for example, economic sanctions. However, in general it should be easier to justify cyber operations on ethical grounds as those operations move away from force on the spectrum of violence.

_Jus in Bello_ – The Law of War

Whereas the _jus ad bellum_ provides a legal framework for determining the lawfulness of a use of force, the _jus in bello_ specifies four principles governing how that force
may be applied during armed conflict. It applies to all parties of the conflict, including the aggressors as well as states operating out of self-defense under Article 51 or in support of a UN operation under Article 39.

Under the *jus in bello*, the legal – and ethical – question regarding a cyber attack is not whether it looks like force, because armed force is permissible, but whether the attack adheres to commonly accepted principles. These principles are embodied in treaties, including Hague Regulations and Geneva Conventions, plus what is called “customary international law.” The latter consists of those practices that are so widely adhered to that they are considered to be legally binding.

The U.S. Department of Defense summarizes the law of war with the following seven principles: (DoD OGC)

1. *Distinction of combatants from noncombatants.* Only members of a nation’s regular armed forces may use force, and they must distinguish themselves and not hide behind civilians or civilian property.

2. *Military necessity.* Targets of attack should make a direct contribution to the war effort or produce a military advantage.

3. *Proportionality.* When attacking a lawful military target, collateral damage to noncombatants and civilian property should be proportionate to military advantage likely to be achieved.

4. *Indiscriminate weapons.* Weapons that cannot be directed with any precision, such as bacteriological weapons, should be avoided.

5. *Superfluous injury.* Weapons that cause catastrophic and untreatable injuries should not be used.

6. *Perfidy.* Protected symbols should not be used to immunize military targets from attack, nor should one feign surrender or issue false reports of cease fires.

7. *Neutrality.* Nations are entitled to immunity from attack if they do not assist either side; otherwise, they become legitimate targets.

The first three principles essentially state that wars are to be conducted by military forces, and that attacks, whether kinetic or cyber, should be aimed at military targets rather than civilian ones. Cyber attacks against critical infrastructures such as civilian energy distribution, telecommunications, transportation, and financial systems would be permitted only if they did not cause unnecessary or disproportionate collateral damage to noncombatants and civilian property.
The first principle also says that military forces should identify themselves when they engage in attacks, thereby taking responsibility for their actions. Part of the motivation for this is so that targets will not blame innocent civilians or other states for attacks and then take actions against them. Applying this to cyberspace, this means that military cyber soldiers should not attack anonymously in a way that leaves open the possibility that they are operating as civilians or on behalf of another state. Because most attacks are conducted so as to avoid attribution, achieving this objective would require novel means and methods, for example, cyber weapons and attacks that carry a government logo or “flag,” or are clearly traceable to a military source. More fundamentally, it would also require a change in perspective, away from the notion that cyber attacks are necessarily covert operations towards one that favors open operations. Governments might oppose this, as it would leave them more open to counter attack.

Although computer intrusions and denial-of-service attacks can be delivered with precision, some cyber weapons could be prohibited on the grounds of being indiscriminate. Most viruses and worms would fall in this category, as they are designed to spread to any vulnerable machine they can find. Viruses and worms might still be used, but they would have to be coded in a way that restricted their spread, say, to a target subnet.

As for cyber weapons causing superfluous injury, there may not be any at this time. However, one could envision a cyber attack that caused such injury, for example, by altering the behavior of a surgical robot during an operation.

There would be ample opportunity for committing perfidy in cyberspace. For example, one could hide Trojan horses on a bogus website that bore the Red Cross logo or place a fake notice of surrender from a wanted terrorist leader on websites used by him to distribute messages. Under the law of war, such acts are not allowed.

The principle of neutrality protects neutral states from attack. To illustrate, suppose an adversary’s cyber attack packets travel through the telecommunications network of a neutral country. It would not be permissible to attack that network to stop the attack as long as the services are offered impartially to both sides and the neutral country is doing nothing more than relaying packets without regard to their content. On the other hand, if the adversary penetrated computers in the neutral country and used them to launch its strike, it would be permissible to launch a counter attack against those machines if the neutral country refused or was unable to help.

In general, then, cyber attacks against an adversary during war could be considered ethical if they follow the above principles. Indeed, they may be less destructive than many kinetic attacks, and thereby preferred on humanitarian grounds. Rather than dropping bombs on a computing center in order to shut down a particular service, thereby causing extensive property damage and possibly loss of life, one might
instead penetrate or disrupt the computer systems in a way that accomplishes the same military objectives but with fewer damages and long-term side effects.

Cyber Attacks by Non-State Actors

Although the law of information conflict concerns state actors and the application of armed force, its general principles can be applied to non-state actors who conduct cyber attacks for political and social reasons. This domain of conflict includes hacktivism, which is the convergence of hacking with activism and civil disobedience, and cyberterrorism, which uses hacking as a means of terrorism. In both cases, the objective is change of a political or social nature, but whereas the activist generally avoids causing physical injury or property damage, the terrorist seeks to kill and destroy.

To apply the international law of armed conflict to this domain, recall that the *jus ad bellum* specifies what types of operations are generally considered illegitimate, namely operations that use force, and the conditions under which these otherwise illegitimate operations can be conducted—conditions that provide a lawful basis for engaging in otherwise prohibited behavior. The *jus in bello*, on the other hand, offers legal principles for the conduct of otherwise illegitimate operations in the face of conflict. The following discusses how each of these applies to hacktivism.

Just Cause for Hacktivism

Just as *jus ad bellum* specifies operations that states are not allowed to initiate against each other during the normal course of events, namely operations that use force, domestic laws specify operations that non-state actors are not allowed to conduct. In the United States, the laws governing cyber attacks are embodied primarily in Title 18, Section 1030 of the U.S. Code (at the federal level) and in state computer crime laws. These laws generally prohibit most cyber attacks, including denial-of-service attacks, web defacements, network intrusions, and the use of malicious code such as viruses, worms, and Trojan horses.

*Jus ad bellum* allows states to engage in otherwise illegitimate operations that use force in order to defend themselves or, under the auspices of the UN, other states that are threatened. Domestic legal doctrine also incorporates a notion of self-defense that allows victims to use force that otherwise would be unlawful. Since the use of cyber attacks as a means of self-defense is covered later, this section focuses on other conditions that might provide ethical grounds for politically and socially-minded hackers to engage in cyber attacks.

One area where hacktivism may be morally justified is civil disobedience, which is the active refusal to obey certain laws and demands of a government through nonviolent means. Civil disobedience is conducted to protest and draw attention to
laws, policies, and practices that are considered unjust or unethical. It employs such means as peaceful demonstrations, blockades, sit-ins, and trespass. Civil disobedience involves breaking laws, but it is an area where violating a law does not necessarily imply immoral behavior. When Rosa Parks refused to give up her seat on the bus, she committed an act of civil disobedience that was morally permissible as well as courageous. Acts of civil disobedience are not necessarily ethical, however. For example, it would be unethical to block the entrance to a hospital emergency room in order to protest the government’s health care policy.

The concept of civil disobedience was extended to cyberspace in the mid-90s. Stefan Wray, founder of the New York based Electronic Disturbance Theater (EDT), credits the Critical Arts Ensemble, which produced two documents, “Electronic Disturbance” in 1994 and “Electronic Civil Disobedience” in 1996. According to Wray, the Critical Arts Ensemble argued that activists needed to think about how they could apply blockade and trespass in digital and electronic forms. (Wray, 1998)

EDT promoted the application of electronic civil disobedience, mainly through “web sit-ins,” which were viewed as virtual forms of physical sit-ins and blockades. Each sit-in targeted one or more websites at a specified date and time, and was announced in advance in a public forum. To participate, activists would go to a website and select a target. This would cause a Java applet called Flood Net to be downloaded onto their computers and generate traffic against the selected website. Although the traffic generated by a single participant would have little effect on the performance of the target website, when thousands participated, as they did, the combined traffic could disrupt service at the target. EDT initially used their web sit-ins to demonstrate solidarity with the Mexican Zapatistas and protest Mexican and U.S. government policies affecting the Chiapas, but later went on to support numerous other causes. The concept was also picked up by other activists, including the U.K.-based Electrohippies. As web sit-ins became popular, the groups also developed more sophisticated flooding software, including software that could be downloaded in advance and run directly from participant machines, and software that required active involvement on the part of the participant (e.g., moving the mouse around).

To assess the lawfulness of web sit-ins and other forms of hacktivism, Schmitt’s criteria for determining whether a cyber attack resembles the use of force vs softer, more legitimate measures are useful. In the domain of activism, legitimate measures include such things as letter writing campaigns, petitions, lobbying, publications, and speaking out. These forms of protest generally come out low on Schmitt’s criteria. They do not cause damage and hence are not severe. Their effects are not immediate or direct, and they are hard to measure. They are not particularly invasive and are often carried out at a distance (e.g., public writing and speaking). They are presumed legitimate. Finally, they are low on responsibility since they can be performed by anyone.
One justification for following this approach is that there is a class of crimes called “violent crimes” that are singled out for their gravity. These crimes use or threaten to use violent force against their victims, and include murder, rape, robbery, and assault. In addition, the concept of civil disobedience expressly calls for the use of “nonviolent” means. Thus, it seems reasonable to evaluate forms of hacktivism in terms of the degree to which they resemble the application of violent force, which is effectively the same as armed force in the domain of jus ad bellum. An alternative approach would be to compare acts of electronic civil disobedience with physical acts of civil disobedience such as trespass and blockades. However, this begs the question of whether the physical acts themselves are ethical. It would be instructive to use Schmitt’s criteria to assess such physical acts of civil disobedience, but that is beyond the scope of this paper.

Using Schmitt’s criteria, let us consider a web sit-in that is publicly announced in advance, is scheduled to last one hour, and produces a noticeable degradation in service. In terms of severity, it would likely rank low, assuming the target is not providing some critical service. Immediacy, however, would be fairly high, as the effects, if noticed at all, would arise once a critical mass joined the sit-in. Directness would also be high. Although impaired performance at the website could be attributed to network problems or increased interest in material posted on the website, the prior announcement of the sit-in all but rules out other explanations. Invasiveness is moderate, but measurability is high, as it is straightforward to measure the performance degradation at the target. Presumptive legitimacy is low to moderate. Even though it is generally against the law to intentionally disrupt service, the effects produced by any individual participant are neither particularly disruptive nor clearly illegal, and the effects as a whole may be minor (indeed, many sit-ins have produced no noticeable effects). Finally, responsibility is moderate. Although it may be easy to determine the group responsible for organizing the sit-in from the public announcement, it would be difficult to determine individual participants. In sum, one measure is low (severity), two are high (immediacy and measurability), and four are in the middle (directness, invasiveness, presumptive legitimacy, and responsibility). Thus, web sit-ins do not look all that legitimate, falling somewhere between lawful measures and the illegal use of force.

Indeed, their legitimacy has been questioned by other activists. Following the EDT’s sit-ins against the Mexican president’s website in 1998, for example, the Mexican civil rights group Ame la Paz objected, saying that the use of hacking tools was counterproductive and dangerous. Another group, the Cult of the Dead Cow, criticized the Electrohippies for their sit-ins, arguing that they violated their opponents’ rights of free speech and assembly. For their part, the E-Hippies justified their actions on the grounds that they substituted their opponent’s forced deficit of speech with broad debate on the issues. They also attempted to justify a planned web-sit-in as part of their April 2000 “E-Resistance is Fertile” campaign against genetically modified foods by asking visitors to their website to vote on whether to carry out the planned web sit-in. When only 42% voted in support, they cancelled the
action. However, they did not offer this option with other web sit-ins, including a massive three-day sit-in against the World Trade Organization in late 1999 in conjunction with the Seattle protests.

Another form of hacktivism is the web defacement. Although most web defacements are not conducted for political or social reasons, they have become a popular tool of protest, accounting for tens of thousands of digital attacks. Outrage over the Danish cartoons of the prophet Mohammed alone generated almost 3,000 defacements of Danish websites. (Waterman, 2006)

One of the earliest defacements took place in 1996 against the US Department of Justice website. The hackers used the attack to protest the Communications Decency Act (CDA), which made it illegal to make indecent material available to minors on the Internet. The defaced website was re-titled “US Department of Injustice” and displayed the message “this page is in violation of the Communications Decency Act!” It also included pornographic images and information about the First Amendment and the CDA. (Attrition, 1996) By displaying pornographic material on a website accessible to children, the cyber attack violated the act that was considered unjust. Considering that the CDA was subsequently struck down as unconstitutional by the Supreme Court, one might argue that the defacement was a reasonable response. However, the defacement also violated computer crime laws, making it much more difficult to justify.

Most web defacements violate computer crime statutes. Examining them in terms of Schmitt’s criteria, they score high on four: immediacy, directness, measurability, and presumptive legitimacy. Severity may be low, as websites generally can be readily restored from backups, but it could be high if the defacement, for example, causes visitors to the site to use erroneous medical information or give up bank account information, or if it severely undermines confidence in the organization owning the website. Invasiveness is moderate, and responsibility is low, in that few countries claim responsibility for such actions. Not everyone is capable of defacing a website, but there are tens of thousands of hackers who are. In sum, defacements look even less legitimate than web sit-ins, and indeed are scorned by many hacktivists.

Other forms of hacktivism can be examined through Schmitt’s criteria. In general, those actions that violate computer crime statutes come out moderate to high, implying their general illegitimacy laws aside. These include cyber attacks to take down websites that traffic in child pornography, and attacks aimed at exposing – and correcting – security vulnerabilities. Even though the ends may be worthy, the means are questionable at best. Acts that would qualify as cyberterrorism would come out high on severity at the very least.

By comparison, cyber actions that do not violate computer crime laws come out low on Schmitt’s criteria. Examples include E-Hippies’ development and use of software to facilitate letter writing campaigns and Hacktivismo’s development of software for
getting information censored in China past China’s firewalls. These activities are lawful (at least in the United States) and do not resemble the use of force.

Cyber attacks that fall in the middle to upper ranges of Schmitt’s criteria are not necessarily unethical, but they are harder to justify. One factor that might be useful is whether the activist’s objectives could be achieved by lawful means. For example, consider again the defacement protesting the CDA. The hacker could have displayed his message on his own website or, with permission, another party’s website, and doing so would have given it a longer “shelf life.” Defaced sites are rarely up for more than a short time, although they may be mirrored in an archive, as was the case here. The defacement got press coverage that otherwise would have been unlikely, but the criminal act is hard to justify given that civil liberties groups were working hard to overturn the CDA through the courts (as they succeeded in doing). Indeed, the defacement could have undermined the legal efforts by linking the civil liberties objectives to illegal hacking.

Conduct of Hacktivism

The seven principles of *jus in bello* provide guidance for using force and, by extension, for engaging in cyber attacks that resemble force.

The first principle, distinction of combatants from noncombatants, states that only members of a nation’s regular armed forces may use force, and that they must distinguish themselves from civilians and not hide behind civilian shields. This principle would prohibit activists from engaging in any form of cyber attack that resembles force. If we interpret web sit-ins and defacements as something less than force, then they might be allowed, however, we might reasonably as that the activists identify themselves or their sponsoring organization so that any response is not directed at innocent parties, including governments. Indeed, the organizers of EDT used their real names and talked about their philosophy and actions in public forum. The E-Hippies were also fairly open, and both groups openly acknowledged responsibility for the web sit-ins they organized. Although the tens of thousands of people who participated in their sit-ins did not individually identify themselves by name, participation in the sit-in itself implied an affiliation of sorts with the sponsoring organization. Web defacers also identify themselves, although typically by hacker group names and individual aliases that are not explicitly linked to their real names. But the level of identification is sufficient for an observer to see that the action was performed by a particular group of hackers and not a government.

The second principle, military necessity, requires that the amount of force employed not exceed the requirements of a lawful strike against a legitimate target. Given that most web sit-ins are conducted against the government agency or company whose policies are the target of protest, they could be interpreted as being consistent with the objective of avoiding collateral damage. However, there have been exceptions. For example, within their broad mission to help the Mexican Chiapas, EDT conducted a
A web sit-in can be viewed as a relatively mild form of denial-of-service attack that affects its target directly. However, there are other types of DoS attacks that leverage third party computers to amplify their affects. For example, in a distributed denial-of-service (DDoS) attack, thousands of third party computers may be compromised and instructed to attack the target. As the compromised machines serve as a shield to protect the source of attack, this would violate the principle of distinguishing combatants from noncombatants.

Many web defacements have been directed only against the government or organization that was the subject of complaint. For example, the Department of Justice, which was the target of the CDA protest mentioned earlier, had supported and defended the CDA. However, numerous other defacements have been against targets that had little if any direct connection to the grievance. Of the almost 3,000 Danish websites defaced in conjunction with the protest against the Danish newspaper that published the cartoons and the government’s response, most belonged to civilian organizations and companies that had nothing to do with the newspaper or government action. However, the attacks did generate press coverage, in part because of their magnitude, likely drawing greater attention to the complaint than simply defacing one or two government sites would have done. Roberto Preatoni, founder and administrator of Zone-h, which recorded the defacements, said that “This is the biggest, most intense assault” he had ever seen. (Waterman, 2006) In general, hackers might justify their defacements of civilian websites on two grounds: first, because the civilian sites were the only ones they could successfully hack, and second, by hacking more sites, they could generate more publicity.

The principle of proportionality requires that any unintentional but unavoidable injury to noncombatants or damage to their property be proportionate to mission benefits. Returning to the EDT example above, the Frankfurt Stock Exchange reported that it was aware of the protest but believed it had not affected its servers. (Denning, 2001) Hence, the sit-in could be considered proportionate to benefits achieved, which arguably were small. By comparison, DDoS attacks affect potentially thousands of noncombatant computers without necessarily meeting mission objectives any better than a sit-in, which does not harm third party computers. Similarly, web defacements against noncombatant servers produce noticeable effects and take time to repair. Besides removing the vulnerability that was exploited and -restoring the home page, system administrators must check for other damage and remove any backdoors and malicious code left behind by the hackers. It is harder to argue that such defacements are proportionate to the protestors’ gains.
Hacktivists have employed indiscriminate computer viruses and worms to disseminate protest messages. These would violate the general principle of avoiding indiscriminate weapons. However, one of the earliest worms, Worms Against Nuclear Killers (WANK), stayed within the network of NASA, the target of the protest. The protestors objected to the nuclear power unit for the Galileo probe.

There do not appear to be cases of hacktivists causing superfluous injury or violating the principle of perfidy. Cyber criminals, however, have exploited protected symbols, including the Red Cross logo, for financial gain (e.g., through bogus fund raisers).

The principle of neutrality implies that activists should not launch cyber attacks against neutral states or third parties. While sponsoring web sit-ins to protest the Mexican government’s treatment of the Chiapas, the EDT conducted sit-ins against U.S. government sites as well as Mexican ones. However, they justified the U.S. sit-ins on the grounds that U.S. policies supported the Mexican government at the expense of the Chiapas.

Other Ethical Frameworks for Hacktivism

Mark Manion and Abby Goodrum offer five necessary conditions for acts of civil disobedience, and by extension electronic civil disobedience, to be ethically justified. (Manion and Goodman, 2000) They are:

1. No damage done to persons or property
2. Non-violent
3. Not for personal profit
4. Ethical motivation—i.e., the strong conviction that a law is unjust, unfair, or to the extreme detriment of the common good
5. Willingness to accept personal responsibility for outcome of actions

Manion and Goodrum’s analysis of several acts of hacktivism suggests they regard web sit-ins, defacements, and some other forms of ethically motivated cyber attacks to be justifiable. However, their analysis ignores their first condition of no damage. Defacements in particular cause information property damage that is analogous to physical property damage (both require resources to repair).

The overall approach taken by Manion and Goodrum differs substantially from the law of war approach taken in this paper. The first principle of *jus in bello*, which states that combatants distinguish themselves, is similar to their fifth condition of accepting responsibility, but the other six principles of *jus in bello*—necessity, proportionality, indiscriminate or superfluous weapons, perfidy, and neutrality – are left out. Instead, Manion and Goodrum rely mainly on the ethical motivations of the hacktivists, taking an “ends justifies the means” approach, at least as long as the attack does not fall in the domain of cyberterrorism.
Kenneth Himma also offers five conditions that weigh in favor of acts of civil disobedience being ethically justified: (Himma, 2006a)

1. The act is committed openly by properly motivated persons willing to accept responsibility for the act.
2. The position is a plausible one that is, at the very least, in play among open-minded, reasonable persons in the relevant community.
3. Persons committing the act are in possession of a thoughtful justification for both the position and the act.
4. The act does not result in significant damage to the interests of innocent third parties.
5. The act is reasonably calculated to stimulate and advance debate on the issue.

Himma’s conditions are stronger than Manion and Goodrum’s, examining means (condition 4) as well as end objectives. However, although Himma’s fourth condition relates to several *jus in bello* principles, it offers fewer distinctions.

Neither framework appeals to *jus ad bellum* for assessing just cause and comparing cyber attacks with acts of force, which are generally forbidden by state as well as non-state actors. On the other hand, both frameworks offer an additional consideration for determining just cause, namely ethical motivation. Further, Himma goes further and asks that activists provide justification for their position and actions; that the position itself be considered plausible by open minded, reasonable persons in the relevant community; and that the actions be designed to foster debate. Himma’s framework is complementary to the law of war framework offered by this paper.

Active Response and Hack Back

“Hack back” is a form of active response that uses hacking to counter a cyber attack. There are two principal forms. The first involves using invasive tracebacks in order to locate the source of an attack. The second involves striking back at an attacking machine in order to shut it down or at least cause it to stop attacking.

The Doctrine of Self-Defense

At the state level, the doctrine of self-defense is based on *jus ad bellum* and *jus in bello*, which together allow states to use force in self-defense, but constrain how that force is applied.

An analogous legal doctrine of self-defense allows non-state actors to use force in order to protect themselves from imminent bodily harm or, under some circumstances, to protect their property from damage. According to Curtis Karnow, formerly Assistant U.S. Attorney in the Criminal Division, the test is whether:
There is an apparent necessity to use force.

2. The force used is reasonable.

3. The threatened act is unlawful.

The necessity condition requires that there be a good faith subjective, and objectively reasonable, belief that there were no alternatives to the counterstrike. The reasonableness condition requires that the harm produced by the counter attack be proportional to the harm avoided. (Karnow, 2003) Reasonableness would also encompass other principles from *jus in bello*, including neutrality, indiscriminate weapons, superfluous injury, and perfidy, as counterstrikes that violated these principles would seem unreasonable. Karnow observes that while self-defense is a privilege of state rather than federal law, it might protect the defender from prosecution under the federal computer crime statute, which prohibits unauthorized access, on the grounds that self-help provides the requisite authorization. (Karnow, 2003)

Karnow also suggests that the legal doctrine of nuisance could justify a counterstrike against cyber nuisances such as viruses and worms. Under nuisance law, a person affected by a nuisance can, as a last resort, use force or other means of self-help to abate or stop it. (Karnow, 2003)

The doctrine of self-defense does not justify retaliatory strikes that are motivated by revenge or a desire to get even. The response must be necessary to counter the threat. To illustrate, in the midst of the Electrohippies’ three-day web sit-in against the World Trade Organization’s website in 1999, the ISP hosting the WTO site, Connexion, conducted a counterstrike against the Electrohippies’ site. Connexion’s server was configured to retransmit all of the attack packets back to the Electrohippies’ website, from where they had originated, thereby shutting it down. Himma argues that the strike back was retaliatory and unnecessary, as Connexion could have simply dropped the incoming attack packets. (Himma, 2006b) Further, the response had a side effect of motivating the E-hippies to develop sit-in software that could be launched directly from their participants’ computers, without the need to go through a central portal. Arguably, this made it more difficult for victims of future sit-ins to defend themselves, as there is no central source for the attack; indeed, such sit-ins more closely resemble DDoS attacks. As another example, the U.S. Department of Defense engaged in active response against a web sit-in conducted by the EDT in 1998. In their case, they redirected the browsers of participants using the EDT portal to a web page with a hostile applet, which caused the participant’s computers to go into an endless loop trying to reload a document. (Denning, 2001) The counterstrike raised legal and ethical issues (some participants claimed they lost data), and the Department of Defense did not deploy similar measures in response to future sit-ins.

Besides self-defense and retaliation/punishment, Himma considers an ethical principle for active response based on the need to secure a significantly greater common good, which might justify aggressive measures. However, he cautions that
such justification can be problematic because of potential unanticipated side effects. He also argues that persons engaging in active response are morally bound to have sufficient reason to believe they are acting on ethical principles. (Himma 2006)

Hack Back and Force

For both state and non-state actors, the doctrine of self-defense allows the application of force against force and threats of force. In general, offensive operations that use less than force call for responses that use less than force. However, even when the offensive act uses force, defensive responses that use less than force are generally preferred over those that use force. Thus, it is useful to know the extent to which active response resembles force vs. more legitimate means, the latter being easier to justify on ethical grounds.

To determine the degree to which a particular means of active response resembles force, we again turn to Schmitt’s criteria. Consider first an invasive traceback such as the one conducted by Shawn Carpenter in Titan Rain. Carpenter traced an intrusion into Sandia Labs and Department of Defense computers back to a province in China. Although the details have not been made public, for the purpose of analysis, assume he had to hack back through computers that were not directly responsible for the intrusion in order to locate the source, as this is typical in cyber attacks.

In terms of Schmitt’s criteria, severity is low. Indeed, the owners of intermediate machines may not observe any effects or even know of the traceback, especially if they had not noticed the intrusion from China in the first place. Given that the effects could go unnoticed unless and until system logs are examined, it seems reasonable to rate immediacy low as well. Measurability is also low in that there is not much to measure. Directness is low to moderate, as it could be hard to attribute the effects of the intrusive traceback to an active response (vs. some other computer intrusion). Invasiveness is moderate, as in all cyber attacks. Responsibility is also moderate, as some skill is required to conduct an effective traceback, but attribution is difficult.

To assess presumptive legitimacy, we need to know who is conducting the invasive traceback and who owns the machines being hacked. If the traceback is conducted by a state actor against foreign systems, presumptive legitimacy should be low in that the entire operation falls in the domain of foreign intelligence collection, which is generally considered legitimate. If the traceback involves accessing a domestic computer, the state may need additional authorities to access the system. However, if the traceback is conducted by a non-state actor, the operation likely violates computer crime statutes, although the offense may be minor if no sensitive information was downloaded or files damaged. But even if we rate presumptive legitimacy moderate or high, the invasive traceback as a whole looks less like force than the cyber attacks examined earlier in this paper. This is consistent with Himma’s argument that tracebacks are not properly characterized as force. (Himma, 2004)
Next, consider an operation that aims to stop a machine from participating in a DoS attack. Suppose that the attacking machine is not even the source of the attack, but rather a victim itself of an earlier compromise. Finally, suppose that the method of stopping the machine from engaging in the attack involves removing malicious code that had been planted on the machine. Severity is low—indeed, removing the malicious code should improve the state of the machine. Immediacy, however, is high: once the malicious code is deleted, the attack packets stop. Directness is moderate, as the attack packets could stop for other reasons. (e.g., the malicious code could be programmed to only attack for one hour on a particular day). Invasiveness is also moderate. Measurability is high, as the before and after attack packets can be counted. Presumptive legitimacy is high, as it is normally illegal to tamper with other peoples’ machines. Finally, responsibility is moderate. In sum, this operation looks more like force than an intrusive traceback, with at least three criteria (immediacy, measurability, and presumptive legitimacy) ranking high. As a result, it would seem harder to justify on ethical grounds. Even though the attack may appear noble—after all, malicious code is removed—it is also more dangerous. Deleting code can introduce problems, as anyone who has had difficulty uninstalling software has learned the hard way. By comparison, one is less likely to cause damage during traceback.

Conduct of Hack Back

Consider again the traceback operation from the perspective of *jus in bello* and the legal doctrine of self-defense. In both cases, a critical question is whether the traceback is necessary for self-defense. Clearly, the operation itself will not stop the attack. Indeed, the most effective way of stopping most attacks is through improved security. However, traceback may be necessary to find and stop a perpetrator who is exploiting an undetermined vulnerability, as the solution would be unknown. Although the machine could simply be disconnected from the Internet, the effect could be worse than the attack itself, resulting in lost productivity and income. In addition, traceback is necessary to find and then stop the perpetrator from going after other targets and causing greater damage. Further, at the state level, traceback may be necessary to identify the source of foreign intelligence collection against one’s own country. Finding that source may be important for national security.

An alternative to traceback is to hand the problem over to law enforcement, but it may be months before law enforcement can even get to the case, let alone solve it. Further, the perpetrator of the attack may have exploited computers in several countries before eventually attacking a particular target, and getting law enforcement agencies in these countries to all participate in the investigation is challenging at best. Moreover, by the time law enforcement responds, the perpetrator may have conducted additional, more serious attacks that could have been averted with a more timely response. Thus, a reasonable argument can be made that at least in certain circumstances, invasive traceback is necessary for a prompt response.
For similar reasons, a traceback that involves invading computers belonging to a neutral country or organization could be warranted if the neutral party is unable or unwilling to stop its own systems from being exploited in the cyber attack in a timely manner.

With respect to proportionality, the seriousness of the cyber attack must be considered along with whether any collateral damage from the traceback is proportional to the harm averted. Whereas traceback may not be justified to defend against a web defacement, it may be appropriate for locating an intruder who has been penetrating a network and downloading sensitive information for months or surreptitiously tampering with or deleting critical data.

With respect to the principles of indiscriminate weapons, superfluous injury, and perfidy, a traceback operation would seem to be in compliance. However, it would not satisfy the principle of distinguishing combatants from non-combatants if the traceback is conducted surreptitiously with the goal of avoiding detection and attribution. To satisfy the principle, the traceback would have to be conducted openly, ideally with permission.

While the above suggests that invasive tracebacks could be ethically justified in accordance with the principles of self-defense, Himma argued that they are not. (Himma, 2004) He based his conclusion on the grounds that they did nothing to either repel or prevent an attack. He further reasoned that tracebacks can locate the source only in direct attacks staged from the hacker’s machine, and, therefore, are unlikely to achieve the greater good of identifying the culpable parties. His argument assumes that a traceback identifies the source of a particular IP packet, but not necessarily the source of the attack. In a later paper, Himma observed that improvements in traceback technologies that allow source identification could lead to a different conclusion. (Himma, 2006b)

Now, consider the hack back to remove malicious code from the victim machine engaged in the DoS attack. A case for necessity is harder to make, as an alternative course of action would be to notify the owner of the machine of the attack and ask that the machine be taken off the network until the code is repaired. Since most owners would not want to risk being held liable for damages caused by their machines, this approach should be effective, although some effort might be required to determine the machine’s owner or get an ISP to notify the owner of a machine on its network. Another course of action would be to get the machines’ ISP to block the attack packets, which at least would stop the immediate attack.

It is also harder to make a case for satisfying the principle of proportionality, given that the hack back to remove the malicious code could potentially damage the victim machine beyond that already caused by the presence of the code, and the operation has no affect on eliminating the original source of the attack. The perpetrator could find another victim and resume the DoS attack from the new base of operation.
With respect to the principle of neutrality, the hack back is also difficult to justify if the victim machine is in a neutral country or owned by a neutral third party. The alternative of notifying the machine’s owner or ISP would be a better choice.

The hack back does not involve the use of indiscriminate or superfluous weapons. Nor does it involve perfidy. However, unless done openly, it would fail to distinguish combatants from noncombatants. The owner of the victim machine would not know who had hacked the machine. In sum, the hack back to remove code appears less consistent with the doctrine of self-defense than the invasive traceback, and thus harder to justify on moral grounds.

Conclusions

This paper has explored the ethics of cyber attacks in three domains of conflict: cyber warfare at the state level, hacktivism conducted by non-state actors, and active response. It reviewed how the international law of armed conflict has been interpreted to cover cyber actions in the context of state-level conflict, and then showed how the resulting framework can be applied to non-state actors and active response.

The framework requires making two determinations: first, whether a particular cyber attack resembles force, and second, whether the attack follows the principles of the law of war. In general, the less an attack looks like force and the more it adheres to the law of war principles, the easier it is to justify ethically. However, attacks that look like force are generally permissible for defensive purposes, so they cannot be ruled out.

To determine the degree to which a particular cyber attack resembles force, the framework uses criteria identified by Michael Schmitt and promoted by Thomas Wingfield. These criteria were developed to distinguish operations that use armed force from softer, more legitimate forms of influence at the state level.

The framework is not intended as a sole instrument for making ethical judgments, but rather as a starting point based on well-established principles. Others have proposed additional considerations that can inform ethical decision making.

Acknowledgements

I am grateful to Ken Himma, Tom Wingfield, and Matt Bishop for helpful suggestions on earlier versions of this paper.
References


