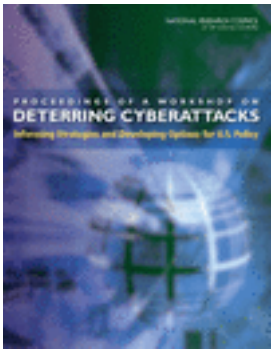

Proceedings of a Workshop on Deterring CyberAttacks: Informing Strategies and Developing Options for U.S. Policy

Committee on Deterring Cyberattacks: Informing Strategies and Developing Options; National Research Council

ISBN: 0-309-16086-3, 400 pages, 8 1/2 x 11, (2010)

Pulling Punches in Cyberspace

Martin Libicki
Rand Corporation

INTRODUCTION

Cyberwar can be considered a campaign that relies exclusively or mostly on operations in cyberspace. Examples might be the 2007 attacks on Estonia or the 2008 attacks on Georgia. States carry it out for strategic advantage. The advantage can be offensive—that is, to change the status quo. It can be defensive—to prevent others from changing the status quo (hence, deterrence et al.). States also carry out cyberwarfare—operations to support primarily physical combat: e.g., purported U.S. operations to disable Serbia's integrated air defense system during the Kosovo campaign.

Yet, a state may find that landing an all-out punch may not be its best strategy. It may, for instance, not convey the message that the state wishes to send. Certain acts of cyberwarfare, for instance, may violate understood norms about the legitimate use of force. Other acts may be escalatory or lead to violence.

Hence the question of this essay: under what circumstance *would* states pull their punches in cyberspace. The argument in this paper falls into five parts. It starts with short parts on the various types of cyberwarfare and a few features that distinguish cyberwarfare from physical warfare. The third part covers norms, specifically the application of existing laws of armed conflict to cyberwarfare and the case for norms more precisely tailored to the conflict in that domain. The fourth part is the paper's core: considerations that states may want to weigh in discussing how far to engage in cyberwar, either for deterrence or for more aggressive purposes. The fifth part addresses command-and-control to ensure that cyber-forces adhere to such limitations.

1 TYPES OF CYBERWARFARE

Cyberwarfare's motivations may be characterized thus:

- *Strategic:* to affect the will and capabilities of opposing states. A hypothetical example might be a Chinese strike on the U.S. infrastructure to warn the United States that the cost of intervening over Taiwan would be felt sharply at home. A strategic cyberattack might be carried out against the main military forces of a state to cripple its capabilities temporarily and allow other states leadtime to prepare for conflict.

- *Deterrence:* to affect the will of other states to carry out attacks, notably but not exclusively, cyberattacks. Cyberattacks carried out in the name of deterrence may be demonstration attacks (although such attacks beg the question of what, in fact, is being demonstrated) and after-the-fact retaliation to convince the attacker to stop attacking, and deter it (as well as onlookers) from contemplating further mischief.
- *Operational:* to affect the conventional (physical) capabilities of opposing states engaged in current hostilities.
- *Special:* to achieve particular effects that are limited in time and place, largely outside the context of physical combat, and usually covertly. Examples may include attempts to hobble a state's nuclear weapons production, to target an individual, or to take down a hostile web site (or corrupt it with malware that others might download). They are analogous to special operations.
- *Active defense:* a potpourri of techniques designed to limit the ability of others to carry out cyberattacks or help characterize and attribute past cyberattacks. Some techniques may straddle the fuzzy line between defense, espionage, and offense.

Note that we specifically *exclude* computer-network exploitation (essentially, spying by stealing information from target systems) and techniques to facilitate computer-network exploitation (except insofar as they disable target systems).

2 WHERE CYBER IS DIFFERENT

The differences between cyberwar(fare) and its physical counterpart are so great that tenets about restraint in the use of physical force are imperfect guides to cyberspace. This part of the paper lays out critical differences between the two. Those interested in a fuller treatment are referred to the author's 2009 monograph, *Cyberdeterrence and Cyberwar* (MG-877-AF, Santa Monica [RAND], 2009).

Here is a summary:

Cyberattacks generally entail the use of information (bytes, messages etc.) to attack information systems, and, typically, by so doing, the information that such a system holds, and potentially affect the decisions made by humans and machines (e.g., power controls).

Cyberattacks are enabled by (1) the exposure[1] of target systems to the rest of the world, coupled with (2) flaws in such systems, which are then exploited. Empirically, systems vary greatly in their susceptibility to cyberattacks and susceptibility may vary over time (especially after systems recover from attack). System owners are typically unaware of all the flaws of their own systems (otherwise they would not be flaws very long).

The *direct* effects of cyberattacks are almost always temporary. Rarely is anything broken, and no one has yet died from a cyberattack (so far as anyone knows). The *indirect* effects can be more persistent: e.g., a target destroyed because a SAM was made to malfunction, a mid-winter power outage in which some people freeze to death.

Cyberattacks are self-depleting. Once a vulnerability has been exposed and deemed consequential, efforts usually follow to eliminate the vulnerability or reduce a system's susceptibility to further such attacks.[2]

The prerequisites of a cyberattack are clever hackers, cheap hardware, some network connection, intelligence on the workings and role of the target system, specific knowledge of the target's vulner-

[1]If one includes insider attacks, then one can talk in terms of the exposure of systems to insiders, although such a broad-brush application of "exposure" does not seem to add much since certain aspects of all systems are routinely exposed to authorized users.

[2]Depletion (of cyber-tricks) could mean that there are only so many tricks and they have been exhausted or the best tricks have been played and what remains (1) produces results that are less useful or easier to recover from before much has been damaged, or (2) works with less likelihood, (3) works under fewer circumstances which are less likely (e.g., the target machine is not in the required state very often). Alternatively the time required to find the next good trick grows steadily longer.

abilities, and tools to build exploits against such vulnerabilities. Cheap hardware possibly aside, none of these can be destroyed by a cyberattack (so, there is no basis for counterforce targeting in cyberwar). Furthermore, none are the exclusive province of states (although states have distinct advantages in acquiring these prerequisites).

Cyberattacks are very hard to attribute. Determining which box the originating attack came from is difficult enough, but even knowing the box does not prove that its owner was responsible, because there are many ways for a hacker to originate an attack from someone else's box. Even finding the specific hacker does not necessarily prove that a state was responsible for his or her actions.

The effects of cyberattacks are hard to measure. This applies even to those directed against well-scoped targets. Systems change constantly: processes that depend on affected systems (collateral damage) are not readily apparent and cannot necessarily be inferred from physical properties. The ultimate cost of, say, a disruption is often directly proportional to the time required to detect, characterize, and reverse its damage; all can vary greatly. Even after a cyberattack, it may not be clear what exactly happened; a data/process corruption attack, for instance, loses much of its force if the target knows exactly what was corrupted. Even disruption attacks, if aimed at processes that are rarely invoked or used as back-ups may not be obvious until well afterwards.

Cyberwar does not sit on top of the escalation ladder, or even very close to the top. Thus, it is not necessarily the last word between states.

3 NORMS

Which of the traditional Western laws of armed conflict apply to cyberwar? Without going through a legal exegesis, suffice it to say that the *technical* characteristics of cyberwar do not allow a particularly clean cross-walk between the laws of armed conflict as they apply in physical space and laws and their application in cyberspace. If norms would apply to offensive cyber operations, there must first be an understanding of the general principles and intentions behind such laws and then rethink their application to the unique circumstances of cyberspace. Some laws will carry over nicely; others will not.

Consider, therefore, the treatment of deception, proportionality, military necessity, and neutrality.

3.1 Deception

The laws of armed conflict frown on making military operators look like civilians, whether among shooters (hence the requirement for uniforms et al) or those being shot (no making command posts look like hospitals). But deception, in general, is the sine qua non of cyberwar. If a message sent to a target system announced "hey, I'm a cyberattack," the target system would doubtlessly keep it out. Cyber defenders take great pains to distinguish legitimate from illegitimate or harmful traffic—this is precisely the purpose of malware protection. Cyber offenders, in turn, take comparable pains to elude these detection mechanisms by masquerading as legitimate traffic. Another form of deception entails making an innocent system or network look interesting as a way of persuading offenders to waste their time rummaging through it, show their cyber techniques to the defender, befuddle them with erroneous information, and perhaps get them to leave the system (falsely) satisfied; such honeypots or honeynets are well-understood but not a common defense tactic.

Should norms frown on making military systems look like civilian systems (e.g., hospitals) in order to persuade offenders to roam elsewhere? The ability to hide looks different in the physical and the cyber world. In the physical world walls and roofs can mask what goes on inside a building—thus indications on the outside can belie what does on inside. In cyberspace, visibility can go all the way through or at least penetrate here and there (e.g., some files are open; others are encrypted). Conversely, in the real world, if walls and floors were invisible, it would be extraordinarily difficult to make the activities of a military command-and-control center look like the activities of a hospital. Absent deep inspection of

content files, it may be difficult to tell what a given organization does simply by learning how it structures its information architecture.

3.2 Proportionality

Proportionality is tricky in both virtual and physical domains. If A hits B and B has grounds to believe that hitting back as hard would not deter subsequent attacks by A, B may conclude that it must hit back harder to convince A to knock it off. In cyberspace the problem of attribution makes overmatch ever more justified: if an attacker can expect to carry out *most* cyberattacks with impunity then the few times attribution is good enough for retaliation may justify striking back hard to make the *expectation of retaliation* an effective deterrent[3] That noted, proportionality is a norm not only because it is just but also because it is prudent if the attacker can counter-retaliate in an escalatory fashion.

Even if the *principle* of proportionality applies to cyberspace as it does physical space, the practical problems in modulating effects in cyberspace to preserve proportionality may be harder. As noted, battle damage may depend on details of the target system that the attacker does not know. Physical attacks at least have the "advantage" of physics and chemistry to work with. Because, say, the blast radius of a thousand-pound bomb is fairly well understood, one can predict what definitely lies outside the blast radius and what definitely lies inside. Error bands in cyberattack are much wider (although some effects can be tailored more precisely: one can corrupt files A and C without worrying about an intermediate file B). Broadly put, the likelihood that a physical attack that exceeds some operational minimum also exceeds some disproportionality point may well be higher in cyberspace than in real space.

The problem of the victim's responsibility about the ultimate damage is an issue in cyberspace, as it is in physical space but much more so. Iraq launched SCUDs against both Iran (in the 1980s) and Israel (in 1991). Far fewer people died per missile launch in Israel, partly because its building standards are better. Notwithstanding whether *any* such terror weapons can be justified, can Iraq be held responsible for the high level of casualties in Iran? Perhaps so because it knew or could have easily known the effects of a missile strike; Iran's building standards should not have come as a surprise. By contrast, matters are more opaque in cyberspace. In theory, well-written software should not fail in ways that break hardware, but poorly written software does exist (otherwise, cyberwar would be impossible). As such, an attack meant to disrupt electricity for a few days (not long enough to harm anyone) may create conditions under which hardware fails unexpectedly, disrupting electricity for months causing incidental deaths. Would it be the attacker's fault if those deaths lead others to judge some retaliation to be disproportionate?

3.3 Military Necessity and Collateral Damage

Can one avoid attacking civilians when seeking to strike the military? Often—especially when military networks are air-gapped, as prudent network management may suggest—but not always, particularly if the target actively seeks to immunize itself by daring the attacker to create collateral damage.

Attackers may have no way to know what services depend on the system being disrupted or corrupted. Some of this applies in the physical world. An attack on a power plant that cuts power to a military facility could also cut power to a civilian facility. One may not know exactly which buildings get their power from which power plants (especially if the target system is prone to cascading failures), but the visible artifacts of power distribution afford a halfway reasonable guess. In cyberspace, neither physics nor economics yield particularly good clues as to which servers satisfy which clients (although sufficient detailed penetration of the server may offer hints of the sort unavailable in the physical world). In an era of "cloud computing" a single server farm may serve any number of different customers, and many may be owned by neutral countries (China's primary airline-reservation systems are located in

[3] *Cf.* Lt. General Alexander's confirmation hearings testimony of April, 2010.

the United States). The problem is not just one of linking a service to its owner; the disruption of an obscure service for a wide variety of customers (e.g., one that reconciles different names into the same identity) can become a newly created bottleneck. The entanglement of services and hence the problem of collateral damage is only growing.

One issue of declining salience comes from replicating malware (e.g., worms and viruses). Although its use may help pry open a lot of doors, these days the preferred attack method is spear phishing—dangling a specific, albeit poisoned, document or link to a malware-infested web site before susceptible users in the hopes that they will bite, thereby introduce malware into their machines, thereby gaining a hacker initial access behind an organization's firewalls.

Avoiding gratuitous harm is a legitimate goal for cyberwar as with physical war, but both depend on some cooperation of the victim (as it does for physical combat). Thus, if the cyberattacker discovers that a particular system exists exclusively for civilian purposes, its disruption or corruption cannot be justified by military necessity (although it may be justified by the requirements of strategic deterrence). This goes double for attacks on systems that affect the personal health and civilian safety. Thus an attack on a dam's control systems that causes it to release too much water and therefore flood a city below it would be considered illegitimate; ditto, for attacking medical files that indicate which medicines go to which people. The target state, correspondingly, has an obligation not to co-mingle systems so that an attack on a legitimate target does damage to protected targets, or at least not co-mingle them more than business logic would otherwise dictate (thus, target states are not obligated to create separate DNS servers for life-critical, civilian, and national security systems).

So, how much knowledge of the attacked systems is required (a problem that applies to kinetic attacks as well)? If the knowledge is deficient and damage results, would opacity on the part of the adversary mitigate the attacker's responsibility? What constitutes a reasonable presumption of connectedness? What constitutes an unreasonable refusal by the attacker to exercise due diligence in examining such connections—does sufficient sophistication to carry out a certain level of cyberattack presupposes sufficient sophistication to determine collateral damage? The warriors-are-acting-in-the-heat-of-battle excuse does not pass the giggle test if warriors are hackers.

3.4 Neutrality

Generally, rules by which to judge neutrals are also more difficult in cyberspace. In the physical world, belligerents ought not cross neutral countries on the way to attack one another. Correspondingly, neutral countries that allow attacks to cross their countries assume a degree of complicity in that act.

But neutrals are not harmed by bad bytes traversing their networks. Can they even detect as much? If they could, then the target should be able to detect and filter them out as well—unless the neutral were (1) more sophisticated than the target, *and* (2) routinely scrubbed malware from traffic that traverses its borders.[4] If so, would it be obligated to report as much, particularly if it jealously guards the existence and capabilities of such a capability? In some cases—for instance, if a fiber optic line crosses a neutral's their territory without going through a local router or switch—it may be incapable of determining the fact of communications, much less their contents.

Regulating behavior *toward* neutrals is also different in cyberspace. In the physical world, country A is not enjoined from bombing a dual-use factory supplying military parts located in country B (with whom it is at war) even if the factory itself is owned by citizens of a neutral country C. Similarly, country A is not enjoined from taking down a server in country B even though it also provides critical services for country C (thus a Russian attack on U.S. servers that hold flight reservations data that could cripple China's air transport system). In practice, the world's information systems are collectively approaching spaghetti status in terms of their interconnections and dependencies. The advent of cloud computing,

[4]Akin to the still-controversial NSA-developed Einstein III programs that are being proposed to carry out deep packet inspection on packets flowing over federal government lines in order to filter out malware.

as noted, only adds complexity. If the threat of cyberwar proves serious, conforming states may have to regard some fraction of cyberspace (e.g., servers, systems, clouds) as off-limits to attack, but this leaves the question of how to assure that no target state uses sanctuary systems to host legitimate war targets (e.g., a logistics database).

If we restrict our consideration of cyberattacks to flooding (DDOS) attacks, some of these distinctions disappear and the laws of armed conflict that apply in the physical world apply somewhat better in the cyber world. Deception is not as important when it is the volume rather than the contents of the packets that creates the problem. The magnitude of the impact, at least the initial impact that can be gauged, and the targets are necessarily specific. But in all of today's fuss over DDOS attacks, they have very limited military utility. DDOS attacks are good at *temporarily* cutting off sites from the general public (and, as such, may affect information operations campaigns, except not necessarily in the desired direction) but militaries tend to be self-contained systems. Designing a communications architecture that mitigates DDOS attacks is not particularly complicated (Google, for one, would be glad to assist for a small fee).

3.5 New Norms for Cyberspace

Norms for cyberwar are better expressed in terms of their effects (what was done) rather than their method (how it was done)—that is, in terms of product rather than process. After all, the faster technology changes, the faster specific processes are retired. Thus a stricture against writing viruses ceases to have much meaning when viruses go out of fashion. Expressing norms in terms of effect are also easier to justify if such effects echo what earlier weapons could have produced; historic resonance helps legitimacy.

All these suggest that whatever norms in cyberspace are developed over the next decades might reflect the spirit of physical norms but they have to assume a different form altogether. So what might such norms look like?

3.5.1 Computer-Network Espionage

One place for norms might be the more acceptable practice of computer-network espionage. At the very least, states should disassociate themselves from criminal or freelance hackers. The practice is strategically deceptive as it permits states to get the benefit of criminal activity without necessarily having to face the international obloquy of whatever such hackers do. Such association also echoes the association between certain governments (notably Iran) and terrorist groups (notably Hamas and Hezbollah); all the major countries consider such associations reprehensible (at least when others do it) and properly so. Such an association is also bad policy: states may be thereby corrupted, may overlook non-hacking crimes carried out by its favored hackers, and may be subject to blackmail (a criminal group under pressure for non-hacker activities could threaten to reveal its links to state-sponsored crimes in cyberspace). It would be even better, of course, if such states took a more active role in prosecuting their own cyber-criminals, but norms against association are a start.

A similar norm might distinguish between national security espionage and all other espionage. The former at least has historical resonance. Spying may even contribute to international stability; U.S. spy satellites circa 1960 collected intelligence on Soviet capabilities thereby assuaging what might otherwise be unwarranted fears that could lead to overreaction. Commercial espionage is simply theft (if intellectual property or identifying information is stolen) or worse (if taking personal information creates the opportunity for blackmail). These norms may be more effectively and appropriately enforced in the commercial realm rather than the strategic realm (e.g., through retaliation or other hostile acts). Thus, by winking at the theft of intellectual property a state might affect its good standing within the international trade community (and violate at least the spirit of trade agreements). That noted, states that think they need CNE for economic development are unlikely to comply or willingly be party to negotiations that

may end up depriving them of such capabilities. Even if they did, enforcement in trade courts (unused to who-done-its) is quite chancy.

A third CNE-related norm might dictate that implants to facilitate subsequent espionage should be distinguishable from implants that facilitate a later attack. Discovery of an implant meant to support subsequent attack may rightfully merit a more hostile reaction than a similar discovery meant to support subsequent espionage. Unfortunately, today's technology may not support the distinction: as a general rule, an implant allows *arbitrary* code to be sent to and run on someone else's system. One would have to engineer implants to forbid certain types of code to be run (including code that would disable the prohibition)—and it is by no means clear how to do that. Incidentally, none of this need mandate that implants in general be easy to find.

This suggests a fourth norm: if an attack on a system is off-limits, then espionage against such a system should be off-limits to the extent that the information acquired from such a system lacks a credible and legitimate purpose. For instance, it is hard to argue that a state must steal information from an organization which is willing to share such information. Although many targets of espionage have valuable intellectual property that they hope to employ overseas, many sensitive systems (e.g., for hospitals, electric power production) belong to organizations unlikely to be harmed if others overseas go about doing their business. Thus, if a state learns of such an implant originated by another state (admittedly, quite hard to determine), the burden of proof that such an implant was for (historically sanctioned) espionage rather than (much more hostile) sabotage should rest on the accused, not the accuser.

3.5.2 Reversibility

One norm appropriate for cyberspace (with little counterpart in physical world) is reversibility: every attack (not intended to break something) would have an antidote and the antidote should be made available to the target when hostilities cease. Thus, an attack which encrypts someone's data (assuming they lack a backup, which no one should, but some still do) should be followed by transfer of a decryption key when peace breaks out. Similarly, an attack that corrupts data should be followed by transfer of the true data (whether this replaced data would be trusted is another issue).

That noted, reversibility is not always necessary. CNE requires no antidote because nothing is broken.[5] Most attacks meant for disruption or even corruption can be reversed by the target's systems administrators in time. In many cases, the corrupted or encrypted data has a short half-life (e.g., the status of spare parts inventories on a particular point in the past) and, for practical purposes, its restoration may be more of a bother. However, this tenet imposes a requirement to refrain from attacks unless there *is* an antidote (this is akin to the rule that those who lay mines have to keep track of them). Thus, under such norms, a corruption attack would not be allowed to randomly alter information unless the true information were stored somewhere else. There may be circumstances where such a requirement is infeasible: storing the pre-corrupted data locally or sending offsite may cue defenders that something strange is going on and there may be no opportunity to ship the data anyhow (e.g., the malware that corrupts the data came from a removable device and the affected system is otherwise air-gapped). Such circumstances will have to be taken into account.

3.5.3 Against Hack-Back Defenses

Certain types of automatic defenses may also be put off-limits. A hack-back defense, for instance, may serve as a deterrent, but attackers who believe such defenses are used may make its attack seem to come from neutral or sensitive site (e.g., a newspaper, an NGO), automatic retaliation against which may create fresh problems. Such a capability is akin to an attractive nuisance (e.g., an uncovered swim-

[5]A separate question is whether the code that permitted or facilitated the CNE should be removed.

ming pool in a neighborhood with children). The automaticity of such an approach is redolent of the infamous Doomsday Machine from *Dr. Strangelove* (albeit with less drastic consequences).

3.6 Practical Matters

None of these tenets lends themselves to easy enforcement. Many of these distinctions are difficult to make technologically. Attackers can deny that they carried out attacks, or deny that the attacks they carried out were designed to have the proscribed effects that were credited to them. In many cases they can blame the targets, either with poor engineering (that inadvertently multiplied the effects of attacks) or, worse, with deliberately manufacturing evidence (e.g., discovered corruption without an antidote) and presenting it to credulous observers. As noted, some of these norms are premature pending the development of technology. The need for norms is vitiated by the upper limits of damage that cyberwar can cause compared to what old-fashioned violence can wreak. Nevertheless, if there are to be norms, these may not be bad places to start.

4 STRATEGIES

This section covers four overlapping issues associated with the management and control of offensive cyber operations: (1) retaliation, (2) *sub rosa* responses to cyberattack, (3) responses that promote de-escalation, and (4) offensive cyber operations in a wartime context. Retaliation will receive disproportionate attention insofar as it discusses matters that may be relevant to other sections.

The issues associated with the management and control of cyberwar parallel those of physical war because they are based on the same political-strategic challenges. However, the manifestation of these issues will look different in cyberspace, sometimes a lot different, and such differences will be explicitly noted.

4.1 Limits on Retaliation

The appropriate level and form of retaliation following a strategic cyberattack (that is, carried out against a nation's economy/society rather than only its military) is governed by several considerations. *First*, is it adequate—does it serve to convey discomfort to the attackers sufficient to make them rethink the logic of cyberwar? This usually sets a lower bound. *Second*, what does it convey about the posture and/or intentions of the retaliator? Such a consideration could ratchet up retaliation (for instance, if the retaliator has a declaratory policy and is worried about credibility), or tamp it down (if the attack appears part of a crisis that threatens to escalate out of control)? *Third*, is retaliation consistent with the recognized norms of conflict, at least as imperfectly translated into cyberspace? This tends to set an upper bound (at least if current Western norms are used). *Fourth*, would retaliation be considered escalatory—how likely is it to persuade attackers to respond with a quantitative or qualitative increase in effect or ferocity? This, too, tends to set an upper bound. *Fifth*, in the unlikely but not impossible event that retaliation is misdirected, what type of retaliation would be easiest to apologize for or even reverse? This consideration is more likely to shape the nature rather than the scope of retaliation.

4.1.1 Adequacy

Although the desire to assure adequacy tends to ratchet the response upwards, it need not do so. Deterrence, it should be remembered, arises from the ability to hit again, not the demonstrated ability to hit the first time—and in cyberspace the two may be quite different. The willingness to pull punches *in the interest of deterrence* is not as paradoxical as it might seem. One reason is that attacks reveal vulnerabilities in the target (and even software or hardware in common use) which are then more likely to be closed once revealed. The more attacks, the harder the target. Thus vulnerabilities disappear. A

limited attack, particularly one that exploits known vulnerabilities unattended by the target's system administrators, reveals fewer new vulnerabilities and thus leaves more undiscovered vulnerabilities (aka zero-day attacks) to hold in reserve as the threat for the next time.

The other reason for restraint is to ensure that the response excites more fear than anger. Any attack may evoke both emotions: fear at an attack's recurrence, and anger at having been hit the first time. But only fear persuades others to do what you want; anger persuades them to do what you do not want.[6] Attacks with too little effect cannot induce enough fear and attacks with too much effect can persuade the target state to respond because it feels that the attacker cannot "get away" with what it did. Finding the sweet spot between fear and anger, if one exists (see Pearl Harbor), is particularly difficult in cyberspace. Not only are power disparities in that medium limited (there is no existential threat, and nations are talented in cyberspace in rough proportion to the extent they are dependent on cyberspace, itself), but the first attack tends to presage a *weaker* rather than *stronger* follow-up attack. Many tricks have been exhausted by being used, hence revealed,[7] and the target may reassess and reduce its current level of exposure to the outside world once it understands the adverse consequences of having been as exposed as it was. However, a relatively light attack that demonstrates capability and resolve (or at least aggression) may hint at the threat of wider and deeper effects without necessarily triggering a level of anger that overwhelms coercive effects.

4.1.2 Consistency with Narrative

Retaliation, particularly if its effects are obvious, makes a statement not only about the wages of sin, so to speak, but also the character of the executioner. This is true in the virtual and physical realm. There will be some nations that will seek to broadcast ferocity; others simply will not care—but the rest, the United States, one hopes, included, will want to make retaliation fit some master narrative about who they are and what kind of rules they would like the world to run by. At the very least, they will want the effects of attack and the means of retaliation to fit some morality play. Certain targets may be regarded as gratuitous and thus off-limits; conversely, if certain sectors of the population (e.g., opposing elites, leaders that hold obnoxious religious views) can be demonized a certain latitude, perhaps vigor, in retribution can be demonstrated. With rare exceptions (e.g., the 1986 raid on Libya) most punitive operations carried out by the United States have been means justified by specific ends: e.g., campaigns against dual-use facilities such as power plants or bridges in order to win wars, air strikes against SAM sites (the 1998 *Desert Fox* campaign against Iraq), or blockades to reduce supplies available to nations that misbehave (e.g., Israel's Gaza policy). Third party hurt was regarded as unfortunate but unavoidable.

What makes retaliation in cyberspace different is that it can rarely be justified by the need to disarm the other side for any length of time. The only justification is to change behaviors and thus harm to third parties such as civilians cannot be excused as instrumental to a tangible military goal. Thus, every reprisal must be judged by the narrative it supports. Such a narrative has to be robust enough to tolerate a wide range of unpredicted outcomes in both directions: fizzles and overkill—both are greater problems in the virtual rather than physical realm. The retaliator may have to pretend that the range of effects produced is a fairly good match for the range of effects sought lest it appear feckless (in promising what it cannot hit) or reckless (in creating more effects than it wanted to), or both at the same time. Precision in messaging is generally unavailable. Nevertheless a certain rough justice may be warranted. If the source of the attack, for instance, comes out of the universities but the state is clearly behind the move, then retaliation that targets financial elites may seem misdirected; conversely, if the attack appears to emerge from organized crime elements, then retaliation that targets the intellectual elite may seem

[6]Unless the point is to anger the target into over-reaction thereby mobilizing the population on behalf of the attacker; terrorism has long tried to exploit this logic.

[7]Although the attacker could have many types of attacks prepared, it presumably led with its best attack—in the sense of most damaging and/or most likely to succeed.

similarly misdirected. Overall, the more that the original attack comes as a surprise to the attacking state's population, the harder it is to justify retaliation that discomfits the same population (unless by doing so, one *intends* to alienate the population, persuade them to support the hackers, and justify the nature of the retaliation in retrospect).

Justification and legitimacy present big potential problems with the retaliation narratives. The attacking state may well deny everything and claim itself to be the victim of aggression (rather than retaliation) unless it chooses to boast of the attack to support a campaign of intimidation against the target state. The retaliator will have to determine how strong a case it can make in public to justify retaliation, but that hardly guarantees that those in the attacking state or friends of the attacking state will necessarily believe it. Very few people understand cyberspace well enough to evaluate the evidence of attribution in an objective manner, even if it were all laid out. Clearer cases in the physical world still beget confusion: after all, the evidence on who carried out the September 11th attacks is essentially accepted in the West, but half of those polled in the Islamic world believe otherwise. South Korea substantiated its claim that the March 2010 sinking of its naval vessel was an attack carried out by North Korea—but it remains to be seen how widely its evidence is believed.

The attacker's counter-narrative also has to be factored in; if it holds that the retaliator is an enemy of the local religion, then it will be reinforced by any cyberattack on religious institutions—hence the dilemma if the retaliator suspects that the original attack *did* come from religious institutions (e.g., that run universities that attackers attend). In the virtual realm, it is not so hard for the attacking state to then attack its own sensitive institution (e.g., a hospital run by religious authorities), demonstrate as much, and blame the supposed retaliator for having done so. Since cyberattacks rarely harm anyone, one moral bar for the attacking government is easy to surmount; after all, it is not causing its own to suffer greatly (indeed, with a little more work, it is possible that one can devise a counter-morality play in which valiant systems administrators stop matters just short of a disaster).[8] Without the kind of physical evidence that might prove exculpatory against such charges (e.g., we did not bomb a mosque because we know the tracks of every flight we sent out and none approached the mosque), the accused state (the retaliator, whether real or supposed) it cannot make its case. All the evidence of who done it is under the control of the attacking state, which can easily argue against giving it up on the grounds that releasing details would show others too much of the system and facilitate future attacks.

Other counter-narratives might take the attacking state off the hook. The state suffering retaliation can deflect the ire of its public against itself if retaliation targets institutions that are not controlled by the state (e.g., banks in some countries). It might do so by blaming the owners for having made themselves open to the schemes of hostile countries by exposing their systems to the outside world without sufficient security (again, this is an argument that almost everyone but the techno-cognoscenti will have to accept or reject on faith). Such a counter-narrative communicates the attacking state's refusal to be intimidated by retaliation, either directly (because it does not yield) or indirectly (because it need not accept the public's blame for the incident). To avoid such an argument, therefore the retaliator may have to find at least some targets that are the responsibility of the attacking state but damage to which is visible to the public (if the latter is the retaliator's intention).

4.1.3 Prosecution Rather Than Retaliation

Unless the attacking country is already at war with the target state (which tends to make retaliation secondary) or boasts of the attack—as it might if the effect was mean to coerce—target countries that profess to rule of law may be constrained to seek justice rather than retribution. Similar issues apply to terrorism, but rarely to conventional war. In the latter case, warfighters presume to act on the state's

[8]In cyberspace, however, accepting that kind of narrative is a matter of faith, with little evidence available to prove or disprove the story—and thus little opportunity for recognizably neutral umpires to call balls and strikes.

behalf, particularly if organized and armed by the state. In cyberspace where organization is a loose concept and being armed even looser, such a presumption may not necessarily apply.

The judicial route buys time. While prosecution is going on, the target state can allay popular ire while the source of the attack is ascertained. In contrast to military narratives that emphasize speed of reaction (re General Patton's aphorism: a good plan, violently executed . . . with vigor now is better than a perfect solution applied ten minutes later), legal processes are expected to be lengthy (the wheels of justice grind slow, but they grind exceeding fine).

Going to court hardly gets the attacking state off the hook. The target state can press the attacking state for access to individuals or information (e.g., let me look at these servers), and use the refusal to cooperate—which is hard to hide—as a justification for retaliation. NATO's invasion of Afghanistan was widely accepted as legitimate even though there is little evidence that the Taliban ordered or even knew about the September 11th attacks—it sufficed that the Taliban sheltered al Qaeda (following the East African embassy bombings) beforehand and refused to turn them over afterwards. UN sanctions were imposed on Libya following the latter's refusal to turn over two suspects in the Lockerbie bombing.

The disadvantages of the judicial route bear note. Many U.S. citizens are nervous about using civilian judicial methods against the leadership of an avowed enemy of the United States (al Qaeda) based, in part on the fear that the accused may escape punishment on a technicality. Arguably, until retaliation *does* ensue, there is no punishment and hence, by some measure, no deterrence. However, there is little to the argument that the failure to retaliate promptly leaves the offending government in place to carry out more mischief; retaliation in cyberspace cannot remove the government by force. Although a retaliation campaign may reduce popular toleration of the government, support for rogue governments tends to rise immediately after it gets into trouble. Governments that lose wars often fall after the war is over (e.g., the Galtieri government in Argentina following the Falklands, the Milosevic government in Yugoslavia following the Kosovo campaign), but this is hardly an argument for instant retaliation.

Conversely, delaying retaliation allows the attacker to bulwark its own cyber defenses so that delayed retaliation has much less of an effect for being postponed—but how much less? First, the attacker, with foreknowledge of possible retaliation, may have bulwarked its own defenses in advance of the attack; only non-governmental systems whose administrators did not get the memo about the upcoming attack would need to use the pause to get ready. Second, the cost and politics of the post-attack bulwarking cannot be ignored. Maintaining a high state of readiness for an extended period of time is costly. For the attacking state to call for systems owners to beef up their cyber defenses in advance of retaliation is to concede to many folks that some retaliation is coming even as it protests its innocence in the original attack. Third, if the two states are mutually hostile, the target may already have implanted malware into the attacking state's critical systems just in case. Implanting attacks do not guarantee that retaliation will work, but it does address much of the problem of gaining access before external barriers go up.

4.1.4 Escalation

Escalation in cyberspace—specifically, the relationship between retaliation and counter-retaliation it—can be a speculative topic. No tit-for-tat exchange in cyberspace has been seen. With no nation arguing that cyberwar is legitimate, few government officials have declared what might constitute thresholds of cyber escalation. The cyber equivalent of Herman Kahn's *On Escalation* is yet unwritten. Not only do we lack a discrete metric for cyberwar, but there is also no good way to measure the proportionality systematically and consistently (e.g., this act is more heinous than that act).

Consider a retaliator weighing between mild and hard retaliation in the physical world. A mild response may fail to convey sufficient displeasure or instill fear; indeed, it may embolden the attacker—who then believes that the retaliator is *unwilling* to strike back hard or *unable* to strike back hard, or both. This is a Munich world. Alternatively, an overly hard retaliation will induce anger in the attacker, shame in that it has lost face by revealing that others have no fear of retaliating against it in a serious way, or a sense of grievance that the punishment was disproportionate to the crime. This is the Guns

of August world. History supports both points of view, but not for the same context. Thus, there is no universal agreement that a harder response is more escalatory than a milder response—let alone what the thresholds are.

Cyberwar adds further complexities. An attacker could be a state, state-sponsored, state-condoned, or just state-located. There is strategic logic in aiming retaliation at a state that did not attack but has the means of thwarting future attacks (e.g., by prosecuting hackers or at least by not harboring and supporting them) so as to persuade them to crack down. But it is unclear whether such retaliation would be considered *legitimate*—especially if the attacking state makes half-hearted attempts to look cooperative (e.g., we'd like to help, but we lack the resources, and letting you nose around in our systems looking for evidence would violate the privacy of our citizens, not to mention our sovereignty). Any non-trivial retaliation might be viewed as unfair, and hence an affront.

The attacking state may conclude it is best off denying everything. If so, non-trivial retaliation indicates that the retaliator thinks the accused attacker is not only a crook but a liar. For the attacking state not to take umbrage at that point is to concede it lied and lose face. It may matter little if the initial protestations of innocence were lies (not everyone will know that it is a lie); indeed, theories of cognitive dissonance suggest that the perceived insult felt by the attacker upon retaliation may be *greater* because the attacker has to work so hard to maintain that it *was* a lie—akin to protesting too much. All this, unfortunately, does not give the analyst much to work with in terms of determining thresholds since the relevant threshold (is retaliation non-trivial) may be quite low (retaliation was always going to be non-trivial). Thus the issue of whether retaliation is met by escalation may have little to do with how damaging the retaliation was; the attacking state will match or, alternatively, exceed the retaliation based on an altogether different set of strategic and political criteria.

Another complicating factor that is far more present in cyberwar than in other forms of response is the uncertainty in forecasting effects. Accidents have fostered escalation in the past (Germany's error that led it to bomb English cities opened the floodgates to city-busting in WWII). They could easily do so in cyberspace, especially if the mode of attack is not system/network disruption (whose effects are often instantaneous and obvious) but data/algorithm corruption (whose effects play out over time and may not be immediately evident to either side without extensive analysis and maybe not even then). If there *were* a recognized threshold in cyberwar, then the presence of great uncertainty in predicting effects argues for doing less rather than doing more as a way of minimizing the likelihood of unwanted escalation. However, the lack of such thresholds means one can be quite precise in predicting effects and quite imprecise in determining the other side's trigger point for escalation. Indeed, the other side may have no clue, or no activation threshold that it can maintain that is not highly context-dependent.

A third factor specific to cyberspace is the relatively narrow band between the least and the most damage one can do by retaliation. Violent war features very wide bands. If your country avoids war with a peer, the number of families that could be suddenly wiped out in any one day is quite low (e.g., from a car bomb or a bad accident). If your country goes to war with a nuclear-armed peer, it is possible that all families could be suddenly wiped out. There is a lot of scope for escalation within that band. Now consider cyberwar. In today's environment, cyberattacks are always taking place (even if acts of deliberate disruption or corruption are relatively rare). Given the number of sophisticated hackers around, one can assume that if a system with requisite vulnerabilities has something worth stealing, theft *will* take place and sooner rather than later. Because the general noise level is high in cyberspace, any retaliation that merits notice as such has to be loud. The top end may also be limited, as well. Who knows how much damage is possible through a no-holds-barred attack on a nation's critical infrastructure? No one has yet been killed in a cyberattack and there is scant indication that a full-blown attack could exceed the effects of a good-sized hurricane.[9] Similarly, there is little revealed information on fail-safe compo-

[9]The most commonly cited worst-case scenario concerns attacks on power companies that damage so much equipment that it would take months to restore power. Yet extracting from Idaho Lab's 2007 Aurora laboratory experiment to such a scenario is quite a stretch—most power equipment defaults to shutting down rather than going haywire.

nents of modern control systems (including for banking) or the extent to which timely and intelligent human intervention can mitigate the effects of cyberattacks. All we know about is what has happened so far, and, by standards of conventional warfare, the damage has not been terribly impressive. Thus the maximum damage may not be so great. All this suggests that the fact more than the level of retaliation will influence a state's counter-retaliation policy if it stays in cyberspace. Conversely, the level of retaliation may influence the original attacker to escalate out of the cyber realm to violence, whether via war or via state-sponsored terrorism.

A fourth complicating factor exists whenever the attacker *wants* retaliation. Terrorists often use attacks to goad the target into responding in an ugly and alienating manner. Fortunately, few states actually use goading as a strategy.[10] The danger is worse in cyberspace. States riven by competing bureaucratic factions may find that one or another faction has attacked the target state in cyberspace as a way of getting the rest of the country to rally behind its particular interests (e.g., a more belligerent stance) or rally against its particular *bête noire* (one bureaucracy may want to take on country A; the other to take on country B).[11] In the physical world, the faction that gets caught may blow its case; in cyberspace it is far easier for a faction to avoid getting caught. Even a state whose top leadership seeks no confrontation can be frustrated by difficulty of enforcing its writ on its own employees. As noted, clever hackers, sufficiently detailed intelligence on the target, and a modicum of hardware suffices to create a capability; thus a faction can, itself, have quite pernicious effects. At that point, the target's response is more likely to play into the hands of the faction that attacked it than into the alternative less-hostile faction—particularly if the suspect faction denies involvement and cites retaliation as proof of the target's/retaliator's malignant nature. Facing an attacker beset with such power struggles, the target state has to ask: would the positive deterrence effect from heavy retaliation trump the negative effect from aligning the state with the faction that identified the target state as the enemy? If the answer is no, the target state may prefer to keep retribution symbolic or even let the incident pass (if the attribution is less than certain, broadcasting doubts can be sufficient reason not to act), respond with judicial means as if the attackers were non-state entities that the state is morally obligated to look for, or try a subtle response (if one exists) which may disproportionately harm the attacking faction vis-à-vis the less hostile faction or the state and its public.

Fortunately, a response does not have to be disproportionate or even equivalent to the original attack to convince the attacker that its calculus of costs and benefits turned negative. Since cyberattacks cannot seize anything directly and only opens a narrow window to permit physical attacks, the tangible gains to the attacker may be modest, and may easily turn negative if the odds of attribution are respectable despite the retaliation being less than overwhelming. A response that is deliberately weaker than the attack may ward off escalation as long as the attacker does not make a case of how baseless *any* retaliation is.

Another approach to escalation asks: what sort of retaliation might put the attacker in a position where it may be forced to counter-retaliate. Clearly, there is a distinction between a cyberattack that kills someone and one that leaves no casualties. Almost as clearly, there is a distinction between a cyberattack with effects made for YouTube and more subtle versions (e.g., a blackout is more visible than a string of ATM failures, not to mention erroneous bank statements). If the attacking regime is nervous about its control over the population, then a cyber-retaliation that affects the performance of its domestic security forces may trigger a panicked and potential escalatory response—but it may also persuade the attacker to back off and argue that bigger stakes than information security are on the table. Systems that the attacker has put public prestige behind—perhaps because they allow it to fill an important promise, or because they have been (mistakenly) touted as secure—may also force the attacker to respond or even

[10]Bismark's Prussia, however, successfully goaded France into foolishly declaring war on it so that it could rally the south German states to its side, beat France, and thereby cement the creation of Germany—but with words, not war.

[11]Japan's army circa 1941 was more interested in combat with China and perhaps Russia; while Japan's navy had its eye on the West's colonies and thus was itching to go after the UK and the United States. The United States in the 1790s found itself divided between factions that favored France and those that favored Britain, each at war with the other.

escalate (without raising the stakes so high that it can be persuaded to call things off). But these are only guesses.

4.1.5 Taking It Back

The last consideration in limiting retaliation is the sneaking worry that one might be wrong. It can affect how one retaliates.

Historically, "oops" is not necessarily followed by "I'm sorry." Even after inspectors gave up looking for WMD, the United States never apologized for having invaded Iraq on false premises. Instead, administration officials argued that Saddam Hussein's failure to "come clean" (cooperate cheerfully with weapons inspectors) left him partially culpable for the West's mistake; it should have behaved as Libya did in renouncing nuclear weapons. False cyber-retaliation may be similarly justified if the supposed attacker failed to welcome the efforts of law enforcement officials with open arms, or did not abjure the capability for cyberattack or the closely related (albeit far better legitimized) capability for cyber espionage. Yet, putting the burden of proof on the supposed attacker to show that it *never* attacked may well be asking for the impossible. However, whereas in Iraq the regime changed (leaving no one to apologize to), such an outcome is far less likely for cyber-retaliation. Thus apologies may be needed to allay a now-hostile state.

Fortunately, it is easier to make amends in cyberwar than it is for physical war. Major destruction is unlikely; mass casualties even less likely. The likely effects of cyberwar are disruption and corruption. Both can be measured in economic terms—and thus if erroneously caused can be partially assuaged by cutting the victim a check. Cyberspace affords the retaliator an even better alternative. Attacks that scramble or corrupt data can often be repaired after the fact by releasing the (cryptographic) keys and thereby let the victim of retaliation unscramble the data or replace the corrupted data. This may not eliminate all the damage—scrambled data may be temporarily unavailable for use, and corrupted data may, conversely, have been used to ill effect—but it helps.

4.1.6 Conclusions to Limits on Retaliation

To reiterate, these five criteria—credibility, communications, norms, escalation control, and reversibility—suggest constraints on the means and extent of retaliation. The first says that retaliation must be noticed but cannot exhaust itself in the process; the next two suggest that it be carried out in a manner that compliments those carrying it out; the fourth suggests avoiding certain types of attacks even if it cannot ascertain a specific tripwire; the fifth suggests the wisdom of avoiding retaliation that is not reversible.

4.2 *Sub Rosa* Cyberattacks

A state's decision to engage only in *sub rosa* cyberattacks means that the attacks cannot awake the target's public and, by extension, everyone else's public. The only states that will know about the attack are the attacker, the target, and those that either side chooses to reveal the attack to. *Sub rosa* options are generally unavailable to attackers in the physical world. Such limitations are meant to ease the pressure on the target to respond by retaliation or escalation.

Table 1 locates *sub rosa* cyberattacks within the context of a broader attack-response matrix. The attacker can choose to carry out an overt attack—which is noticeable and boasted about (e.g., for the purposes of coercion). Alternatively, the attacks can be obvious—noticeable by the public but not owned up to. Finally, the attack can be covert; pains are taken not to make its effects public, and, no claim is made. (The fourth possibility—the attacker gives itself credit for an attack that it is at pains to hide—makes little sense). The retaliator has three choices: retaliate openly, retaliate in an obvious manner, or retaliate covertly.

Table 1

	Attack is Overt	Attack is Obvious	Attack is Covert
Response is Overt	Open cyber-war	Retaliator has to run the risk of explanation in the face of what may be the attacker's denials as well as the risk of possible error.	Retaliator has to reveal the attack (to mobilize its population, perhaps), and then run the risks of their denial.
Response is Obvious	Why bother? No one is fooled.	No one is fooled by the response, and error is possible. Yet, it lets both sides "deny" everything if they tacitly agree to settle.	Revealing the original attack would justify retaliation but point the finger at the retaliator; conversely if the retaliator is caught the *tu quoque* defense looks contrived.
Response is Covert	Puts the onus on the attacker to reveal what happened and explain why. Retaliator may have to answer to public about why no response.	Signals displeasure but also a desire not to let things get out of hand. May not deter third parties (except via rumor). Will also not protect against error.	*Sub rosa* cyberwar. Signals displeasure but also a desire not to let things get out of hand. Third parties know nothing. May protect against error.

Now consider the responder's options by type of attack (Table 1).

If the attack is overt, the responder is being dared to respond. It could take the dare openly. It could carry out an obvious cyberattack and not take credit, but it will be assigned credit anyhow because it has an obvious motive. Alternatively, it could respond covertly, leaving leave the original attacker to decide whether to reveal the retaliation—which, by the nature of a covert attack is likely to target state institutions that do not deal with the public—or keep quiet and boast that it attacked with impunity. The attacker's advantage lies in keeping quiet if interested in forging a tacit or covert agreement to stop. If the attacker does keep quiet, and particularly if it continues attacking, the responder will have to answer to its public (and others) why it did nothing. The responder's primary motive for keeping matters covert may be the hope that it can, in fact, persuade the attacker to stop without making it obvious to others *why* the attacker stopped.

If the attack is obvious but not overt, the primary consideration for the target is how certain it is of who attacked. Overt or obvious responses present similar considerations. Even with obvious responses, few who think about the matter will have serious doubts that the target was, in fact, the retaliator since only the target had an obvious motive.[12] The only reason to deny what is otherwise obvious is to permit both sides to save face ("What, hackers we?") while they come to a *modus vivendi*. Finally, a covert response to an obvious attack signals displeasure but also a desire not to let things get out of hand. It may lead to a tacit or at least covert settlement. The responder could concede that it had been attacked but claim it had insufficient facts to warrant a counterattack (and hope the target of its counterattack keeps quiet). A covert response will, however, not deter third parties, and it will not protect against error (see later in this section for why).

If the attack is covert, the responder has a deeper dilemma. It can respond overtly in order to make an example of the attacker—who is likely to deny its culpability and may even demand proof that such an attack took place. Such a strategy may be pursued to mobilize opinion against the attacker, particularly if the original covert cyberattack is a prelude to overt hostilities. An obvious response may be chosen because the target list with an obvious response is better. If the attacker reveals the attack, its doing so

[12] There are exceptions: (a) the attacker has struck multiple countries and so the retaliator can be one of several countries, (b) the attacker has multiple enemies even if they all have not been struck in cyberspace, or (c) the attack gives a third party who may dislike either the attacker or the target an opportunity to weigh in. In the last two cases, the real counter-attacker has to have something in wait ready to be sprung at just the right time. Nevertheless, the exceptions are exceptional.

will suggest to others who the true author of the response is. The risk is that if the responder is fingered and then claims it was attacked first, albeit covertly, such an argument will appear contrived.

The purest case is where both sides attack each other covertly. The attacker may wish to exert pressure on the target state's leadership without causing a public reaction that may constrain the ability of the leadership to act in the attacker's interest (or least not act against the attacker's interest). The retaliator may wish to dissuade further attacks without itself riling the attacker's public. In other words, both sides are likely to believe that each can come to an agreement and avoid escalation if each side's public is not, in effect, consulted on the matter. In the unlikely event that both side's leadership consist of hawks afraid that their publics are dovish, they can carry on at each other without undue interference (the less likely scenario inasmuch as having outsider observers tends to make concessions harder lest leaders lose face). And so the game goes on until someone concedes either explicitly or tacitly or until one or the other side's attacks are made public—either through one side's decision to reveal all or part of the exchange, or because what was supposed to be covert turns out to be overt. Alternatively, the exchange may continue indefinitely until the target systems have so hardened themselves that attacks are no longer worthwhile.

The retaliator may also wish to limit itself to covert responses because of attribution problems. If it is confident that it knows who the attacker is, but cannot or will not provide a convincing rationale to the rest of the world, then a covert response puts the onus on the target to explain why it is being attacked. But a covert response has a sneaky way of indemnifying the retaliator against the consequences of the retaliator's errors. If the retaliator is correct, the attacker will probably have a good idea who hit back because the attacker knows who it hit (unless the attacker was overly generous when selecting targets). If the retaliator is incorrect, however, the unfortunate victim of retaliation may be left hurt but confused: it does not know about the original attack and therefore has no reason to suspect the retaliator (to be sure, other evidence may reveal who the retaliator is, so a covert response is not entirely risk-free).

Sub rosa cyberwar can be quite tempting for the responder and even the attacker particular within the covert operations or covert intelligence community. No one has to produce evidence of attribution, and there is less pressure to reveal the particulars—methodologies and targets—of the attacks. Indeed, it is far easier, in most respects, to carry out rogue operations in cyberspace than it is to do so in the physical world: e.g., fewer prisoners to worry about. Unfortunately, what is most attractive to some becomes a weakness to others. The covert community lacks the public oversight that more overt parts of the national security community operate under. If it wishes to justify its actions, it has more control over what evidence is collected and presented; it has less to fear from contradictory material from neutral or hostile parties. Members of the covert community, despite their personal probity and honesty, tend to operate in a sealed world. Mistakes can go uncorrected for longer. When actions are criticized, there is a reflexive tendency to circle the wagons. Even those who argue that members of *our* covert community are like the rest of us, only in different professions, the same may not hold for members of *their* covert community where rule-of-law is noticeably weaker.

The second problem with *sub rosa* warfare is that each side's strategy is hostage to the exercise of discretion by the other side (not to mention the accidental revelation of covert attacks). Once revelations start everyone may end up looking bad: not only the attackers on both sides, but also the targets who could be called both feckless (for having been had) and sneaky (for covering it up). On the one hand, a primary rationale for keeping matters covert is to facilitate later settlement; on the other hand, those in the know—that is to say, the covert community—generally do not populate the sort of organizations whose primary motive is to reach accommodation with the other. Covert communities, by their nature, distrust all other covert communities. So, each side has to weigh whether it is better off pulling back the shades on these *sub rosa* exchanges or letting matters continue their subterranean course. The result may be a game of chicken. Each knows that revelation will make its side look bad not only to the public, but perhaps also to its own masters, but each may hope that the *threat* of revelation may make the other side look worse. Each side, may therefore be in a position to concede things to hide their common activities in ways that might be impossible were their "negotiations" subject to public scrutiny. The dangers seem obvious.

As noted, a covert cyberattack must abjure certain targets (i.e., a limit on offensive cyberwar options). Clearly, targets that affect the broad population are out; thus, attacks on critical infrastructures are incompatible with a *sub rosa* strategy. Perhaps, one can imagine an attack that could be and *would be* characterized as an accident but reported as deliberate (thereby achieving its dissuasive impact) by those who operate the infrastructure—but large conspiracies of this type are heir to leakage and hence failure. The same would hold for attacks on financial systems—unless bankers, too, want to play that game. That leaves as primary targets, parts of the government that do not interface with the general public (or whose interfaces can be controlled so that glitches in the interaction are made good or otherwise covered up). Still, those who run systems that were hacked have to be counted on to keep quiet about what they have seen. Ironically, this leads to two observations. First, that the best targets of *sub rosa* cyberattacks are systems associated with the covert community. The very best targets may be systems that the target state is reluctant to admit having.[13] That noted, intelligence systems tend to be air-gapped and thus essentially unavailable for attack. Second, as an open society, the United States is a poor target for a *sub rosa* attack because of the difficulty of keeping attacks secret. Closed societies offer good targets for *sub rosa* cyberattacks because they keep secrets better. This further implies, incidentally, that states at war offer more *sub rosa* targets than those at peace because wartime conditions tend to reduce the amount of information that traverse the boundary between a military and the rest of the world.

4.3 Promoting De-Escalation in a Crisis

How does a crisis begin?

War in the physical realm generally requires a concert of movements (e.g., tanks moving from garrison, leaves cancelled, civilians mobilized, supplies surging to the front) which must take place so many hours and days before conflict if forces are to be ready when war starts. No such physical or economic constraints spell the onset of conflict in cyberspace. Even attacks that have to be prepared, perhaps years in advance, cost only a little to maintain (e.g., to check if the intelligence is still good and if the implants are still functioning and accessible); they can be postponed for weeks and months often with little harm. Might a good indicator of an attack be an acceleration of bulwarking actions as the attacker braces against an expected retaliation? Perhaps, yet many such preparations can be made near-instantly if planned properly.

4.3.1 A Phony Crisis Can Start with Finding Implants

Indications of cyberwar are more likely to tell where than when (analogously, finding troops marching to your border says when, but finding them hidden in the hills says where). Current practice favors planting code into targeted systems well in advance of any crisis. An implant is designed to activate a subsequent strike; it allows the attacker to send arbitrary attack code to the target system confident that the target system will attempt to run the code at privilege levels which can cause serious damage to the system's functions or integrity.

Finding their attack code in your system may constitute a crisis—but what kind? As with an attack, attribution is more difficult if no attack has occurred. First, because an implant is not a time-specific preparation, there may be no real-world crisis going on at the time of discovery to suggest who may be planning a cyberattack. Those looking for what earlier crisis prompted the implant may be frustrated by the fact that intermediate servers that would allow one to trace the original infections may have long ago purged their memories. Even if the memories exist, one cannot know who did it if unable to trace back months' worth of packet traffic, a near-impossible task. Second, implants need not contain attack code,

[13]In the wake of the controversy over DARPA's Total Information Awareness program, funding was ended. If, as many believe, the program went underground into the intelligence community, those who run such systems may be quite reluctant to admit that they exist.

just enough code to permit attacks to be attached to operating system or common application. Without attack code to analyze whatever forensic clues exist in the code to trace it back to its putative writer are simply missing. The one advantage to finding an implant is that many attacks erase themselves after completion leaving nothing to analyze, but this is true only if they are completely successful and are not placed on machines that studiously archive everything to ineradicable media.

Last is the problem of distinguishing an implant prefatory to an attack from an implant prefatory to taking information. Most implants are in the latter category: they exist to keep the door open for the next attack or to facilitate the process of raising the privilege level of the rogue code to whatever is necessary to carry out its spying mission. Discovering an implant proves that the bad guys got in, but it does not prove that the attacked system can be harmed apart from having its files read.[14] It gives no clue to the likely success of the attack or its consequences.

So what does the target of an implant do? Presumably, the discovery prompts it to check its systems for similar implants in particular, or rescrub systems for implants in general (now that evidence of one suggests the possibility of others). If the state wishes its private and non-federal sector (or even its own less attentive bureaucrats) to scrub with alacrity, it may have to raise alarms about what it has managed to find. (Until these implants are vacuumed up, any emergency cyber-defense short of shutting off affected networks from the rest of the world may be meaningless; the enemy, so to speak, is already inside.)

Need the target state accuse specific others and thereby start an *external* crisis to allow it to meet an *internal* (clean-up) crisis?[15] In other words, need it make the case (tendentiously under the circumstances) that it has found a particular enemy's implant rather than simply an implant in general? Perhaps yes. It may want a crisis and this discovery furnishes as good a pretext as any. Similarly, it may wish to justify retaliation in another domain (e.g., to weaken the opponent). It may wish to embarrass an opponent, characterizing it as hostile (for having made the implant) and incompetent (for having let it be discovered) at the same time. In such circumstances, forbearance is secondary—the target state wants a crisis for its own purposes. Perhaps no. The target state may honestly be concerned that the supposed attacker should have been but was not been deterred from making an implantation. In today's world in which CNE is expected, the accusing state has to credibly argue that the implant could *only* have been meant for an attack. Not only is it unclear what would constitute evidence, but finding enough people who could and would evaluate such an argument on its merits may also be daunting—more daunting than, say, similar arguments about aluminum tubes made just prior to the 2003 invasion of Iraq.

In general whatever crisis follows finding an implant is one of choice not necessity.

4.3.2 A Real Crisis Needs an Attack

A nation that has been attacked, but has yet to respond and restore the *status quo ante bellum* can be said to be in a cyber-crisis. In common with most crises, a cyber-crisis has an internal component (managing domestic expectations) and an external component (managing international expectations and preserving one's "face" abroad).

In both cases the internal and external crises sharply change their character if and when there is a consensus on who attacked. Because people tend to blame others, notably specific others, for their problems, an internal consensus is likely to form first—indeed almost instantly (albeit not necessarily reliably) in some cases. States reluctant to escalate crises (which is most of them most of the time) may find themselves buffeted by internal popular pressures to respond and great uncertainty about whether that might start a completely unnecessary fight.

[14]The notion that any process that can read a file can also alter it is misleading—the process may not have sufficient write privileges, or the applications that read the data may ignore data that is not digitally signed, and the likelihood that a rogue process can replicate digitally signed data without access to the signing key is infinitesimal.

[15]At least one, "senior American military source said that if any country were found to be planting logic bombs on the grid, it would provoke the equivalent of the Cuban missile crisis." From *Economist*, July 3, 2010, "Briefing: Cyberwar," p. 28.

Until there is a *global* consensus on who attacked, the target can do nothing and not lose face *overseas*. Once such a consensus emerges—and it may emerge even if the technical analysis of the attack is quite incomplete—the target state will be viewed as having been challenged, as if slapped by a metaphorical glove in a duel. The challenge may be overt—attribution could be conceded and the attack justified as a response to a cyberattack, to the need to pre-empt a cyberattack (if it feels the target audience is unsophisticated and unaware that pre-emption is nearly impossible), a festering wound, to non-cyber forms aggression, or to mounting hostility from the target. The attacker may deny the attack directly but praise (and support) the attackers (e.g., the attitude of Iran towards Hezbollah). Or, the attacking state may deny everything with a wink and a sneer—as if it wanted to insulate itself from the automatic application of legal norms but wants the rest of the world to know that it has the capability to carry out such an attack and the target lacks the capability to ward it. Alternatively, the attacker may have been surprised to have been accused so soon, and turn sharply about-face and admit is culpability in order to wring whatever it can out of the target's humiliation (the 20th century is full of crises where the aggressor sought to humiliate others).

The target state, operating under the eyes of the world, can evaluate the attack in many ways: error (the attacker thought, erroneously, it was hit first), inadvertence (e.g., the attack was an accidental or unsanctioned act carried out by a rogue faction or at the behest of the attacking government), culminating (e.g., the act was undertaken to right a past wrong, or as a warning shot to defuse an existing crisis), or provocative (the act was a signal or prefatory to more hostile action). In the first three cases, the pressures on the target state emanate from its own citizens and from those in the rest of the world who may believe that the attack was provocative. The attacker, itself, may not view forebearance negatively but as deserved in the first case (error), a blessing in the second case (inadvertence), and a statement of maturity in the third case (culminating). Only in the last case (provocative) will it find a failure to respond to be an act of weakness that deserves further exploitation. The question for the attacker in the first three cases is whether it can provide enough political cover to the target to convince third parties that the failure to respond harshly cannot be considered a sign of the target's weakness but its strength in being able to see past recent pain to achieve more long-term goals. Or, the question may be moot: chances are the attacker (or supposed attacker) was accused because it was not on good terms with the target state in the first place.

In great contrast to physical war, the *direct security* implications of doing nothing are benign. A nation that does not respond to an invasion may cease to exist. One that does not respond to a raid, however, may continue. Cyberattacks are more like raids whose effects are temporary. Since retaliation cannot disarm the attacker, the failure to react has no bearing on the attacker's ability to do damage.

How a state's response appears to third parties has less to do with which category it comes in. The issue of whether the target should be embarrassed—and thus look vulnerable and perhaps pathetic—for having been fallen hard as a result of an attack remains, irrespective of the attacker's motive. This is not something that strike back will fix. Indeed, broadcasting a message of "how dare you expose my weakness" exposes it all the more by revealing the target's sensitivity to pain. As for deterrence, unless the attacker helps out (notably in the case of inadvertence or culmination)—apologizing or at least explaining the attack rather than boasting about it or denying it—motive may not matter terribly much either. This may put the target in an uncomfortable position if its interests vis-à-vis the attacker call for *not* retaliating (or not doing so in any but a nominal manner)—one reason to avoid declaring a deterrence olicy.

How should the target state distinguish the first three cases (where forbearance may be advisable) from the fourth case (where forbearance can be misinterpreted as weakness)? Forensics can help distinguish bad from good attribution, but its ability to distinguish inadvertence from deliberate attack is limited to its ability to distinguish one attacker from another within the same state, a weak clue about intention at best. One test of intentions may come from putting pressure on the attacking state to separate itself from the actual attacker; this is a test the attacker can pass (carrying out an open coercion campaign after conceding the rogue attacker seems illogical), but not necessarily fail—a lot depends

on prior relationships (is the attacking state in thrall to the attacker's group?) or subsequent relationships (have the hackers been made heroes?). Finally distinguishing between a culminating attack and a prefatory attack depends almost entirely on political clues. Hints—but hardly proof—that the attack was culminating include statements to that effect, offers to reduce tensions, a visible turning to other matters, or a stand-down of forces (although how one can determine that cyber forces are standing down is a real puzzle).

4.3.3 Self-Restraints for the Attacking State

What should attackers, themselves, do to keep tensions in check?

Ironically, some of the question applies to falsely accused attackers. Unfairly or not, they may have some convincing to do. Simple denials may not communicate very much—indeed they are expected. Conclusive evidence that some named third party carried it out may be viewed with suspicion until verified. Offers of assistance with recovery may be viewed askance since it is difficult to help without getting privileged access to systems of the sort that the attacker would pay dearly to obtain. Unfortunately, the same fog of cyberwar that makes it difficult to prove who did it, also make it difficult to prove who did not do it. The least bad approach may be to give the target enough access to satisfy itself—unless the target is being driven by ulterior motives.

If escalation is to be avoided, a real attacker should make the case that it seeks no wider war; the attack in question was narrowly tailored in both scope and time, and not designed to permanently change relationships or power (it helps that the effects of cyberattack are usually temporary). If cyberattacks are used as punishment and nothing more, there may be value in being explicit and even verifiable (at least if facts matter to the target). *Thus, if an attacker would avoid escalation, it may help to put certain targets off limits so that its motives are not confused.*

One class of targets to avoid is systems whose malfunction can endanger others: e.g., health care, water supply, standalone safety systems (e.g., 911), system-specific safety systems (e.g., dam controls), and traffic management (unless declared explicitly in advance—e.g., do not fly after midnight). Such attacks are generally contrary to the laws of armed conflict, anyway.

Attacks that might threaten permanent changes in state power may contradict a narrative of limited effect. These include attacks that look prefatory to military action such as those on strategic warning systems, or more broadly, such as corruption attacks that may make the target worry about the quality of its command-and-control (including weapons C2). A related category may be morale-related attacks on armed forces on the basis that changes to morale are not restored when systems that support them are restored.

Similarly, crippling systems that hamper the ability of the target state to maintain its hold on power maybe misinterpreted as prefatory to a regime change campaign. To be avoided are disruptive but especially corrupting attacks on state-friendly media and internal security systems. Something like the Great Firewall of China would be off-limits as well, despite how richly apropos a target it may appear. A related set of attacks to avoid are those that undermine the basic trust that citizens have in their government and comparable institutions: e.g., corruption attacks on the financial system.

So, what *would* be allowed—particularly if trying to impress a regime that only cares about the risk of permanent loss? Against such targets, the prospects for containing a crisis may be weak—but then the use of cyberattacks to get its attention may be pointless. Otherwise, if dealing with a state that pursues limited ends, and is willing to weigh the benefits of its counter-attacking you against the further costs you can impose on it, there are many ways to convey displeasure, and tilt its calculus without running afoul of your narrative. They include disruption attacks on infrastructures (even financial infrastructures) as long as they are not aimed at causing much permanent damage or put lives at risk, disruption attacks on government agencies both civilian and military, and corruption attacks on both domestic and international intelligence activities (just enough to force the adversary to institute new more tedious procedures to ensure the quality of its information, but not enough to make it feel naked in its ignorance).

To reiterate a point made above, do not assume precision, or anything like it, in gauging the effectiveness of such attacks, or even the certainty with which attacks do or do not cross some line.

4.3.4 Forbearance in Wartime

The calculus of forbearance looks different in wartime. When blood flows, combating states will find that cyberattacks are the least of their reciprocal problems. Decisions elsewhere will overshadow their efforts at signaling one another in cyberspace. If, however, they are involved in testing one another (e.g., Egypt and Israel during their war of attrition), both sides may *want* to test their strength in cyberspace, believing full well that each can gauge the other side's capacity to give and take punishment without sparking the other to drive toward its capital (the tendency of each side to harden targets under conditions of cyberwar may echo similar but slower developments in other domains).

Those who limit their shooting to air, space, naval, or marine domains (offshore islands, remote possessions) because they fear uncontrolled escalation may similarly want to contain cyberwar. At the very least they would avoid targets whose disruption can only be justified by their effect on the national morale and the popular willingness to support further war. Military attacks directly associated with such a conflict (e.g., against air-defense radars) are presumably fair game and would not be seen as specifically escalatory. Even if cyberattacks precede violence, by the time both sides react physical force may already have been used (if the use of force is delayed too long the effect of the cyberattacks may have been reversed).

Can either side carry out cyberattacks against civilian targets to influence the fighting, *without* each touching the other's strategic nerves? Although similar issues arise over physical attacks (e.g., striking ports that support naval operations), they can often be dealt with through geographical limitations on combat (e.g., no bombs north of the Yalu River circa 1951). Boundaries in cyberspace are harder to define and confine. The reported U.S. strike on a jihadist web site[16] supposedly took out 300 servers around the world. Indeed, the information processing (e.g., C4ISR) support for combat operations generally need not be anywhere near the conflict (RF bandwidth permitting) and are more survivable if they are not (a subtle adversary may deliberately outsource such processing to cloud farms of third party countries within which one encrypted database is indistinguishable from another). Thus, the useful boundaries have to be logical rather than physical ones. Unfortunately, as Tom Schelling points out, for such boundaries to be effective in limiting the activities of both sides, they either have to be negotiated or so obvious as to suggest themselves (e.g., stopping at the river's edge). Otherwise they seem arbitrary and meaningless, or concocted to favor the side that advocates them. The nuclear threshold was one such boundary. The distinction between fatal and nonfatal cyberattacks may be another.

Asymmetries between opponents will complicate tacit agreements on what to leave intact in the cyber world, just as it does in the physical world. A local conflict between the United States and China over Taiwan will take place much closer to China: agreeing that homeland ports are off-limits favors them (they have no reasonable prospect of attacking debarking ports in California); the reverse favors the United States. One country may use coal to generate its electricity; the other, hydropower. A policy that has each side refrain, for safety reasons, from interfering with dam controls unfairly penalizes the coal-using state whose electrical generating capacity alone remains at risk. States that have built dedicated communications lines for defense are disadvantaged against states that must depend on dual-use infrastructures if both agree not to target dual-use nodes (routers and switches). Countries that feed intelligence to "patriotic" hackers to carry out cyberattacks are at an advantage over those who depend on their own employees if the onus against cyberattacks is levied on the basis of standard command-and-control dicta.

Attacking military facilities outside the field of war is often frustrated by the likelihood that they are air-gapped. Yet, many legitimate targets of cyberwar have dual uses—are they off-limits? Should one off-limits from a physical attack be off-limits from a cyberattack *that offers the potential of similar*

[16]Page A1 of the *Washington Post*, 19 March 2010.

collateral damage? Thus, if ports are off-limits for fear of harming civilian dockworkers, should cyberattackers refrain from corrupting its databases for fear of disturbing commercial operations overly much? In reality, attackers may now know what they damaged. They cannot count on help from information managers, reluctant to be seen as responsible for having poor cyber-defenses (leaving the ultimate victims, say shippers, wondering why schedules slipped)—but whose reticence to speak at least works against escalation.

Basing mutual forbearance on nuance and precision cannot be serious. Differences in judgments about who did it and what they did will persist. Third parties may also muddy the waters, particularly if they are interested in undercutting mutual forbearance. Apart from not wandering too closely to self-proclaimed limits, what can either side do to make its intent clear? Defenders can distinguish attacks by military opponents from others by reasoning that their foes have no interest in wasting their assets (knowledge of the opponents' vulnerabilities) on low-impact attacks—but only if they so choose. But little prevents the opponent from looking like (let alone making arrangements with) third parties for the same ends. The fog of war is an argument against forbearance.

This leaves a question as relevant in cyberspace as the physical world: if such gentlemen were so capable of negotiating such nuanced norms, why are they still resorting to fighting to settle their differences?

4.4 Proxy Wars

Another situation in which cyberwarriors may have to pull their punches is when going after the systems of a state which is helping an active combatant.

Local wars offer great powers the opportunity to contribute to the fight (and to upset their peers) by supplying information (e.g., about the battlefield or other information systems), information services, and cyberwar services to indigenous forces—and with seemingly little risk. Nevertheless, if proxy wars are not to escalate into general wars some boundaries on cyberattacks have to be set. In physical combat—using the Korean and Vietnam wars as examples—the bounds between allowable and proscribed targets were iffy but mostly observed. Chinese forces were fair game for U.S. forces below but not above the Yalu River. Russians avoided the Korean theater except for (rumors of) air combat. No one attacked U.S. forces out of theater. During the French-Indochina War, the United States was liberal in sending supplies, but not people. In the Vietnam War, similar rules applied: in theory Russian and Chinese "advisers" to North Vietnamese forces manning Russian or Chinese equipment (mostly SAMs) were not explicitly off-limits, but some U.S. policy makers worried about unintentionally killing them (while others were disappointed that they were rarely hurt).

The extent to which information systems of one great power will be considered off-limits to cyberwarriors of another great power (backing the other indigenous force) may depend on whether such help is considered akin to offering supplies (hence, largely protected) or offering forces (hence, less protected). The fact that information system assistance tends to involve activity on the part of (cyber) warfighters says forces, but the expected immunity of cyberwarriors from harm says supplies. The help offered by the great power by way of mapping opposing networks, revealing their vulnerabilities, and crafting exploit tools may be so complete, that the additional effort of sending commands to activate an attack would seem almost an afterthought—the fact that the great power does not pull the trigger may not reduce its culpability appreciably.

Even if both great powers agree to target only those systems in indigenous hands, the links between indigenous combatants and the systems of their great power friends may influence whether these great powers end up going after each other's systems. Can indigenous systems, in fact, be attacked without interfering with systems of their friends? Are the great powers' in-theater systems connected to its global systems? If one great power harms another great power's systems, would the target state want to make an issue of it? Can the attacker argue that such an attack was essential to helping its indigenous ally? Can the target retort that the attack was really meant to harm its systems and not indigenous systems?

Absent the right kind of firebreaks, one can imagine a continually escalating confrontation that requires either negotiations to establish mutual limits on the spot, or for one great power to back down unilaterally lest general war in cyberspace ensue.

So what norms should apply? One possibility is look for whatever opportunities exist to use physical boundaries as proxies for cyber boundaries. Thus, if the friend's systems are outside the war zone, perhaps they should be off-limits to a cyberattack even if they help the indigenous ally fight (as supplies destined for the war zone would be if they were sitting in the territory of the great power friend). This begs the question of how the attacker would know where any such systems sit. It also leaves the problem that cyberattacks against the indigenous combatant's systems may affect the great power's systems, something the attacker may have no way of knowing beforehand. A cross-infection may be *prima facie* indication that the two systems are not, in fact, virtually separated.

Asymmetries plague the application of such tenets in practice. For instance, the boundaries between systems of the great power and its indigenous friend may be well-defined and guarded on one side but not the other. Thus, the escalation risks from attacking the former would be low but the escalation risks from attacks on the latter would be high. Why should the more careless side get a free ride just because the attacks of the more careful side have the potential of riskier outcomes? Worse might ensue if attacks on the indigenous combatant's infrastructure end up bedeviling the lives of citizens of its great power friend (incidentally, such systems could easily sit in third countries).

Avoiding escalation in such scenarios might require proxy warriors to carefully separate their global systems from those sent to theater and require attackers to exercise great caution to ensure that their cyberattacks have precise effects. But it would not hurt for either external force to realize that accidents happen, especially in war zones.

5 IMPLEMENTATION

States that would pull punches in cyberspace must have appropriate command and control of their cyberwarriors. Instructions on what to avoid must be clear and the controls must be in place to ensure that such instructions are followed.

In the physical world, both command and control are getting better thanks to ever-more-ubiquitous surveillance and the proliferation of communications nets (e.g., cell phones). The effects of war can be meticulously documented and attributed. As more military equipment becomes digitized (and thus capable of hosting copious log files), the prospect of knowing exactly who did what when draws closer.

Not so in the cyberworld, where keystrokes can come from anywhere. Standard operating procedure (which is anyway rather thin for cyberwar) is a poor guide when one cannot state *a priori* exactly what the means of attack are (for instance, the route in often has to be determined in real time as the contours of the target's defenses are revealed) much less what the effects of attacks are. Any policy designed to attack up to some boundary but no farther is subject to two uncertainties: the difference between what was intended and what actually takes place, and the difference between what takes place and what is perceived to take place. If the odds and the cost of escalation are significant, the only possible lesson may be to stand as far away from the border as possible: in other words, do nothing in that domain. If one would act, clear and *thick* margins of some sort have to be established.

The burden of margin-setting will differ depending on whether one is worried about, alternatively, careful, careless, and rogue cyberwarriors.

Careful cyberwarriors are those that pay as much attention to constraints as they do to results. For them, clarity is the goal. The constraints on their behavior could include how to attack, and what results are unacceptable under which circumstances. They should be explicit, advertised, and somewhat stable (or at least not fluctuate arbitrarily). The rules that say what actions are permissible in what situations should be codified in advance of crisis because when the fighting starts, purposes are more fluid, and not necessarily broadcast to all (also true for physical combat). To make constraints work, it may be neces-

sary to teach the basic principles of cyberwar as they apply to national security. Beyond such guidelines, the rules on how to attack or what constitutes non-excessive damage may be too context-specific to be specified too far in advance.

Careless cyberwarriors mean to follow the rules, but in the heat of combat, may convince themselves that carrying out a clear operational mission trumps conformance with often-ambiguous guidelines. All the rules for careful cyberwarriors apply to careless ones (who may be indistinguishable from careful warriors). The application may vary: the actions of careless warriors are apt to drift over the borders, and, being human, likely to blame their trespasses on unclear guidance, the ambiguities of cyberspace, and even the target's behavior (e.g., turning off the electric power substation to disable government bureaus was not supposed to put hospital patients at risk; where were the latter's backup generators?). If careless cyberwarriors are a problem, one approach would be to limit the amount of intelligence *all* cyberwarriors are provided with. To wit, if a particular target is likely to be off-limits, then probing such targets is neither resourced nor tolerated—but what if no target is not off-limits in some contexts? If so, there may be no target that cannot be probed for its susceptibility to cyberattack. Once such intelligence is gathered on such targets, they may be vulnerable to the efforts of careless cyberwarriors even if they are off-limits in that particular context. Postponing collection until war starts (and thus its context is clear) is problematic if collection becomes that much harder, afterwards. Such dilemmas have echoes in the physical world. Japanese F-15 fighters were designed not to have drop tanks, which thus limited their range, so as to demonstrate to the Soviet Union that Japan posed no *offensive* threat to Soviet territory.

The last category contains rogue warriors—so eager to strike the target that they take their work home with them, sometimes literally. Trained and filled with intelligence at work, they carry out attacks from platforms or intermediate conduits that are very difficult to trace and out of sight of their supervisors. Rogue warriors will not respond to constraints when freelancing (except as hints as to what to avoid appearing to do). Because they do not have to work in military formations or with unique military hardware, they are harder to detect and hence control than their equivalents in physical combat: e.g., the militias of developing nations. Not even keeping them (figuratively) chained to their desk in a crisis will eliminate mischief if they have found how to contact their own bots from their desktop—although such behavior may be suppressed if they have to account for every keystroke (unless the rogue operator creates malware that goes off unless specifically told not to). After all, even careful cyberwarriors are unlikely to carry out mischief from a ".mil" network lest the target of their attentions filter out their rogue packets simply by looking at their return address. Effective militaries have ways of filtering out most such rogue warriors and engineering social controls that keep potential rogue warriors in the force from straying. Having done what they can, states then have to determine whether the risks of violating self-imposed constraints merit reducing every cyberwarrior's access to the intelligence and tools necessary to mount the more sophisticated attacks.

6 CONCLUSIONS

Cyberwar is a messy, messy, messy, messy business. It is messy by the very nature of cyberspace where it is nearly impossible to know that the connections you cannot see do not, in fact exist. It is messy because the instruments of war are not necessarily monopolies of states, much less states in the immediate area of conflict. It is messy because knowledge of what happens to or what ensues from one's enemies is opaque and subject to manipulation. Finally, it is messy because one's own command-and-control is far less certain than for physical warfare. These, properly understood, create enough reasons to be skeptical about the instincts to declare that cyberwar unavoidable in one or another context.

In confrontations that are less than existential—and cyberwar, standing alone, clearly fits that description—an important role is played by the correct alignment of actions and narratives. The latter combines an interpretation of the relevant context of action in terms of a state's moral (and concomitant legal) structures. The adherence to norms, in turn, can be expression of the narrative, at least for states that take norms seriously rather than rhetorically.

The question of pulling punches gains another meaning in the context of war, but war may not necessarily be the best metaphor for the deliberate mischief that takes place in cyberspace. In many ways the metaphor of pollution is a better descriptor. In other contexts, cyberattacks may be said to be what happen if security engineering receives insufficient attention (just as accidents are what happen if safety engineering receives insufficient attention). A boat with a leak will take on water as fast in friendly waters as it does in hostile waters; the leak matters, the waters do not.