Deterring Cyber Coercion:
The Exaggerated Problem of Attribution

David Blagden
Strategy and Security Institute, Department of Politics
University of Exeter
d.w.blagden@exeter.ac.uk

This is an open-access draft of an article accepted for publication in *Survival*:
It represents the version that the journal accepted for publication, but has not undergone copyediting, production, or post-acceptance updates

Accepted: 2 January 2020

Abstract
Can cyberattacks be deterred? As the capability and associated damage potential of cyber weapons rises, this question will become ever more important to publics and their policymakers around the world. The principal obstacle to deterring cyberattacks via the threat of retaliatory punishment is usually taken to be such attacks' ability to be made technically untraceable: absent a 'return address' for the aggression suffered, how could the victim of such an anonymous attack know where to direct its retaliation? Such concerns are overblown, however, for they conflate two distinct variables within the deterrence calculus: aggressor *identity* and aggressor *interests*. In fact, once cyberattack is understood as the coercive political act that it is, the 'anonymity problem' for cyber deterrence dwindles. This is because, in seeking to advance a cause via cyber coercion, an attacker must necessarily reveal a set of interests that it values. Such interests can then be held at risk by the party seeking deterrence, even if the attacker's identity itself remains concealed.

Acknowledgements
I thank Ben Buchanan, Andrew Futter, Erik Gartzke, Jon Lindsay, Kubo Mačák, Patrick Porter, Simon Smith, Peter Watkins, *Survival*'s editors and reviewers, and especially Helena Mills for invaluable comments on drafts of this article. Of course, all errors and omissions remain my own.

Can cyberattacks be deterred? As the potential damage threshold of possible future cyberattacks rises in line with the sophistication of cyber weapons,[1] this question is increasingly salient for governments, national security agencies, and the populations they seek to protect[2] – especially as concern grows over the probable centrality of cyberattack to future hostile 'hybrid' operations.[3] This article explains why one characteristic of cyberattack, its potential for anonymity, is not the insurmountable barrier to effective deterrence based on retaliatory punishment that it may seem. For coercion – defined as cost-imposition in the pursuit of behavioural change – is the form of hostile political influence that deterrence seeks to oppose. And in order to coerce, a belligerent must necessarily identify that which it values, i.e. the interests it seeks to advance via coercion. Attempted coercion, in short, serves as a preference-revelation mechanism. Thus, even *if* a belligerent *can* escape identifying itself via anonymous cyberattack, it can*not* escape identifying interests that it holds dear, which can then be held at risk by those seeking deterrence.

This argument builds upon an important pre-existing research insight, namely that the political interests being advanced by a cyberattack will often make the identity of the aggressor clear, even if the origin of the attack itself cannot be readily traced via technical means.[4] This represents the difference between 'strategic' and 'tactical' (i.e. technical) attribution: while a cyberattack may be *technically* anonymous, the *strategic* interaction of which it forms a part can remove much of the presumed anonymity, by making clear whose interests are at stake. Yet that still leaves open the concern that many actors could have similar interests, meaning that the specific aggressor might remain unidentified – so even for analysts who have identified the importance of strategic attribution, it remains a circumscribed hope.

This article demonstrates, however, that such reservations over strategic attribution are overblown. This is because they conflate two different variables within the deterrence calculus: aggressor *identity* and aggressor *interests*. The two variables do often correlate, of course: if you want to retaliate against an attacker's territory (the interest), say, then you need to know *whose* territory (the identity). Yet the linkage is not necessary: it is possible to identify valued interests of an attacker without identifying those interests' sole holder. And crucially, coercion is not possible without identifying the interests being advanced, in cyberspace or anywhere else, for the very nature of coercion requires revelation of desired changes in behaviour.[5] If you are Russia, for example, you do not need to identify *which* specific NATO country cyberattacked you to still identify – and hold at risk – interests shared by *all* NATO members. It is therefore possible to retaliate against revealed preferences – to

achieve successful strategic attribution of *interests* – even when strategic attribution of a specific attacker *identity* remains challenging. Many actors may indeed hold similar interests while only one of them may be responsible for a particular cyberattack, but that need not necessarily matter: the very variable that obscures their specific identities, i.e. their shared interests, also means that retaliating against those interests will punish the underlying aggressor. The fact that the specific attacker may have 'got away with it', in terms of not being identified individually, will be little consolation when their interests have nonetheless been retarded. And knowing as much *ex ante*, it is possible for deterrence to hold based on the threat of *ex post* retaliatory punishment against 'anonymous' aggressors.

Why study the deterrence of cyber coercion? After all, cyber weapons are ill-suited tools for explicit coercive signalling, since their capability is often (a) unclear prior to their use and (b) degraded by their revelation (i.e. defenders can patch/circumvent the vulnerabilities that the threatener intended to exploit).[6] Many categories of hostile cyber action – attrition of an opponent's capabilities, hacking for enjoyment/publicity, covert cyber-espionage, cyber-enabled theft for private/state financial gain, subversion/disinformation, and so forth – are also pressing security concerns, but do not achieve their effects via the issuance of coercive threats. And it is not as if cyber-deterrence has been an unambiguous success thus far; there have been plenty of cyberattacks against states with retaliatory means at their disposal, yet which (for whatever reason) failed to generate successful deterrence. Several high-profile attacks against US and allied corporations – such as 2012's attacks on the US financial system and Saudi Aramco, 2013's attack on the Sands Casino, and 2014's attack on Sony Pictures – have gone undeterred, for example, despite successful culprit attribution *and* America's superior relative power (vis-à-vis Iran and North Korea respectively). On top of all that, one could question the value of analysing this topic in principle rather than via case studies, since theoretical understanding of deterrence is already well-developed while empirical scholarship on cyber-aggression remains (relatively) nascent.

None of these rejoinders holds water, however. By themselves, cyber-threats are indeed flawed tools of coercive signalling. But taken in their strategic context, a wide range of hostile cyber actions meet a reasonable definition of coercion: cost imposition in pursuit of behavioural change.[7] If an attacker seeks to coercively exploit a favourably changed balance of power following successful attrition and/or extortion, if a protestor draws attention to their cause in a way that generates coercive political costs, or if cyber-espionage delivers information that is utilized coercively in another domain, then there are coercive interests being advanced.[8] Observed cyberattacks, meanwhile – attacks that, for whatever reason, have

not been deterred – are not evidence of the futility of the deterrent enterprise, since defenders' resolve to impose punishment rises with the severity of attack suffered while the pool of capable-enough potential culprits shrinks. So, North Korea going undeterred in its 2014 Sony hack, for example – a rudimentary and low-damage attack, all told – does not preclude that deterrence is vital to understanding the absence of great-power usage of (more capable) cyber-weaponry against other major powers.[9] And the value of a conceptual treatment[10] lies in theory's ability to impose clarity for both scholars and policymakers alike on a question that – due to the unavoidable selection bias that afflicts studies of the causes of non-occurrences – will necessarily under-reflect deterrence's causal significance when studied empirically.[11]

The article first discusses how the risk of escalation to levels of retaliation that carry costs higher than an aggressor could bear can deter aggression at lower levels of violence, but also how the anonymity of cyberattack could compromise such an escalatory deterrent ladder. Second, the relationship between coercion and deterrence is explored to show that each is necessarily tied to the promotion and defence of political interests, and thus that attempting to coerce via cyberattack without revealing one's preferences is a contradiction in terms. As such, even where the specific identity of an aggressor is not revealed by the attack itself, there are nonetheless revealed *interests* that can be held at risk by an aspiring deterrer. Third, several caveats to the core argument are laid out.

For reference, a cyberattack is understood as the intentional use of computer code – a 'cyber weapon' – to harm or exert hostile control over another party's information and communications technology (ICT) systems or networks, and/or the physical systems and living beings dependent upon them.[12] 'Cyber deterrence' is understood as the act of deterring potentially hostile cyber operations; this may or may not be conducted using the deterrer's *own* cyber capabilities.[13]

The article concludes that the anonymity of cyberattacks need not prove a barrier to effective deterrence via established methods of escalation linkage between the initial attack and subsequent, more costly levels of retaliatory punishment. That said, this approach carries downsides: it is premised upon targeted retaliation against interests some group of attackers value, which could include 'innocent' states and/or civilians – thereby risking both moral opprobrium and political blowback – plus the interaction of cyber capabilities with cross-domain (conventional/nuclear) retaliation may generate unstable escalation. As such, the finding does not imply that efforts to harden ICT systems should be neglected, since such denial capabilities can nullify the need for retaliatory punishment.

Escalation Dominance, Deterrence, and the Problem of the 'Return Address'

Deterrence is achieved when an actor – a state, an armed group, or an individual – concludes that the likely costs of an attack exceed the likely benefits.[14] Opponents can be deterred in two ways.[15] Deterrence by denial occurs when a potential aggressor concludes that attack will be ineffective because of the strength of the countermeasures in place: it is *denied* the opportunity of achieving its coercive objectives. Deterrence by punishment, by contrast, does not diminish an attack's effectiveness, but rests on the threat of retaliation of sufficient magnitude that the aggressor will be worse off than if they had not attacked in the first place: they will be *punished* after the fact. Deterring cyberattack via effective denial measures, i.e. sufficient ICT hardening/resilience that attacks are not attempted – or fail if they are – is of course an important component of effective cyber-deterrence. But the focus here is deterrence of those threats that cannot be reliably denied, for which the threat of retaliatory punishment is central to deterrence hopes.

There is no reason in principle why the use of any new weapon cannot be deterred via the threat of retaliation using the same or alternative weapons systems. To be genuinely beyond such 'cross-domain' deterrence – deterrence reliant on the threat of cost-imposition in an operational domain other than that in which the original attack occurred – a new weapon would have to possess destructive capability that no other existing weapon could match *and* be available only to one of a pair of enemies. And while the latter may indeed be the case for certain cyber weapons, given differential rates of capability development, the former will never happen for cyber weapons of *any* capability, because nuclear weapons can already impose essentially unlimited costs.[16] Richard Clarke's complaint that cyber deterrence via the threat of retaliation lacks credibility because the capabilities of potentially-retaliatory cyber weapons are kept secret[17] – or would be nullified if revealed – thereby omits that cyber weapons are only one potential retaliatory option.[18]

Of course, nobody is suggesting retaliation against irritant-level cyber disruption using, say, thermonuclear genocide – and even if they were, such a posture would struggle for credibility, not least because of the humanitarian consequences' moral/diplomatic costs. The point, however, is that the risk of escalation to more serious retaliation – with the counter-attacker's own cyber weapons, conventional forces or, ultimately, nuclear weapons – raises the potential costs of even low-level attacks. With the advent of nuclear weapons, a plausible escalatory pathway exists whereby the costs of attempted coercion can *always* be made to exceed the expected benefits – a variant of what is dubbed 'escalation dominance' in

deterrence theory.[19] None of this is to suggest that such cross-domain escalation is desirable.[20] It is precisely because it is so *un*desirable that the danger of triggering such escalation is sufficient to deter potential aggressors, even if both sides are aware that the worst-case end-point is unlikely.[21] This logic underpinned Cold War 'tripwire' military deployments; U.S., British, and French garrisons positioned in West Berlin, for example, could not hold the city against the Red Army, but created an unacceptable risk of escalation to general war with NATO's nuclear powers if Soviet forces attacked.[22]

Reflecting the attraction of deterrence by punishment as a buttress to attempted denial, its prominence in Western cyber-strategy – led by the United States and United Kingdom, the two most capable Western cyber-powers[23] – has grown in line with capabilities and understanding. Neither the US nor UK 2011 cyber defence strategies mentioned retaliation once, for example, although both referred to deterrence and dwelt extensively on hardening and resilience-building (i.e. denial).[24] Following Edward Snowden's revelations of Anglo-American capabilities, however, both have become less reticent to publicise their full range of options. The 2015 US cyber defence strategy, for instance, stressed that 'The United States must be able to declare or display effective response [i.e. retaliatory] capabilities to deter an adversary from initiating an attack,' while the 2016 UK version similarly has an explicit sub-section on offensive capabilities as a component of deterrence.[25] As part of this, the Western allies have also made it increasingly explicit that 'cross-domain' retaliation – i.e. retaliation against cyberattack using non-cyber capabilities, if the severity of the harm merits it – is a component of their thinking, in a bid to generate the escalation equivalence necessary to sustain a credible punishment posture.[26] Most recently, twin US posture innovations that follow from the 2018 National Cyber Strategy – the 'Cyber Deterrence Initiative' (which aims to bolster retaliatory deterrence through collective alliance responses) and the 'Persistent Engagement' doctrine (which strives for a continuous cycle of tracking and offensive action against emerging cyber threats) – display yet more emphasis on punishment logics.[27] There are important questions, of course, about whether such growing faith in the efficacy of offensive operations – especially of the pre-emptive variety – as a response to threatened aggression could itself become a source of escalatory danger, as discussed below.[28] The point at this stage, however, is that just as escalation to the imposition of ever-greater costs has been an integral component of past deterrent postures, so too it is already an important component of major powers' responses to emerging cyber threats – and unlikely to become less so.

For effective deterrence via the threat of punishment, however, a potential attacker needs to know there is a significant chance of actually being punished. Assessments of this likelihood focus on the defender's capability and resolve: does the victim of an attack have the incentives, determination, and wherewithal to retaliate by imposing a sufficient level of costs?[29] Yet embedded in the 'capability' half of the equation is a crucial sub-question: can the aspiring deterrer *target* their retaliation in the first place? Absent a 'return address', no amount of retaliatory firepower will deter if attackers can be confident it will never be directed against anything they value.

This is where cyberattack poses a challenge for deterrence. For as noted, sophisticated cyberattacks may prove untraceable, even by the most advanced cyber security agencies.[30] It also seems unlikely that technological innovation will suddenly make all future cyberattacks traceable.[31] The victim of an attack possessing cyber, conventional, and even nuclear forces could therefore be irrelevant if they have no idea where to direct such firepower: a potential attacker could expect to act with impunity, resulting in deterrence failure.[32]

It is beyond the scope of this article to contribute to technical debate over the attribution of cyberattack. Nonetheless, via techniques of internet protocol (IP) address masking, routing attacks via numerous connected computers ('bots') that have – wittingly or otherwise – been turned into a cross-border network, initiating an attack from a cybercafé or public library, hacking and utilizing some unsuspecting individual's internet-connected mobile phone, and so forth, cyberattackers may be able to conduct aggression against internet-connected targets without significant risk of their location or affiliation being revealed.[33] Furthermore, even if the computer used to initiate an attack *can* be an identified, it is a greater challenge still to prove who was sitting at it, or on whose orders they acted.[34] The ongoing state-versus-private arms-race in encryption technology adds to this difficulty.[35] Hacking non-internet-connected ('air-gapped') ICT systems is more challenging – often requiring a 'real-world' intelligence operation to gain physical access to a server, say – but that too can be achieved, given sufficient capabilities, with the potential to avoid attacker identification.[36]

In short, cyberattacks may appear immune to the retaliatory deterrence pathway described previously. This article now demonstrates, however, that cyberattacks are not actually as anonymous as their technical characteristics imply, and so do not enjoy the immunity from targeted retaliation that is often feared.

Identities versus Interests: Punishment and the Coercive Revelation of Preferences

Key research in cyber deterrence and coercion is united around the pivotal importance of attribution. For Thomas Rid and Ben Buchanan, 'Attribution is fundamental: almost any response to a specific offence – law enforcement, diplomatic, or military – requires identifying the offender first'.[37] Erica Borghard and Shawn Lonergan similarly contend that 'coercion in cyberspace requires attribution to be effective', while Martin Libicki reasons that coercion must be visibly associated with the coercer to operate, the logic being that the coercer must identify themselves if they are to cause the behavioural change they desire.[38] In this they are aligned with seminal early unpackings of the role of punishment in deterrence, which simply treated the existence of an identified enemy against which retaliation could be directed as a baseline assumption.[39]

Those early renderings lack precision, however – and in doing so, they obscure a crucial distinction in the logic of punishment, with substantial implications for the cyber-deterrence debate. Retaliatory punishment is about imposing costs on *interests* – political, economic, and social 'goods' that the target-actor *values*, be they publics, cities, forces, wealth, status, or anything else from which decision-makers derive utility. The specific *identity* of the attacker, by contrast, is not actually an interest – it is a mere assignation, with no necessary political content. Of course, many of the very highest interests are exclusively *associated* with a particular state identity; the 'mere assignation' comes with a bundle of *ascribed* political content, tied to the interests with which it is synonymous. So, if one wishes to achieve deterrence by threatening some package of political, economic, and social 'goods' that are located in a particular territory, identifying the owner of that territory becomes central to the pursuit of deterrence. This explains why the elision of interests and identities was largely unproblematic in the deterrence literature of Cold War bipolarity – for most practical purposes, the interests that each aspiring deterrer sought to hold at risk were synonymous with an exclusive identified adversary. Put simply, the continued existence of the people/communities/industries/conurbations/ideologies/etc that constituted the United States and Soviet Union were the unique and pre-eminent interests of two specifically identifiable actors named 'United States' and 'Soviet Union'.

The elision of interests with the identities of those who may hold them has been a recurrent source of analytical missteps in the debate over the deterrence of cyberattack, however. Rid and Buchanan are correct to argue that 'attribution is fundamental', on one level – but this does not by itself tell us what is being attributed (i.e. interests, identities, or both). Many interests are indeed unique to a specific holder – but many others are not.

Indeed, for that very reason, Borghard and Lonergan's contention that 'coercion in cyberspace requires attribution to be effective' is circumspect. For if we understand cyber-coercion broadly as cost-imposition in the pursuit of behavioural change using cyber capabilities, rather than solely explicit threat-issuance – as discussed earlier – then there are plenty of conceivable coercive cyber-activities that serve the interests of multiple actors, without any one culpable actor identifying themselves.

The crucial point for the feasibility of deterrence based on the threat of retaliatory punishment is that one does not *need* attribution of a specific attacker identity, *provided* that one can attribute relevant interests. If an aspiring deterrer can identify some aggressor 'good' on which unacceptable costs can be imposed – along with possessing both the capability and resolve to actually impose such costs, in the potential aggressor's estimation, as with all deterrence calculations – then there is no reason why deterrence via the threat of retaliatory punishment cannot hold, even in the absence of specific-attacker identification. Deterrence by punishment has always been about threatening to harm aggressors' interests rather than their identities *per se*, in cyberspace and everywhere else; it simply happens that identifying a sole holder of certain interests makes that task easier in many circumstances. So if the anonymity of cyberspace makes it undeniably hard to reliably attribute specific attackers – technical/forensic attribution may be possible, but sometimes will not – what prospects does it offer for the identification of threatenable interests?

Just as war is an extension of politics by violent means,[40] so too deterrence – the threat of imposing unacceptable costs on an aggressor by causing their failure in war (denial) or via retaliatory war (punishment) – is necessarily a political act.[41] In deterring attack, a state (or non-state group) promotes intrinsically political interests: the continuing peaceful and prosperous existence of a human community, its social preferences and values, and the political entity they have created and inhabit. The flip-side of deterrence is coercion: using force or the threat of force to impose costs on an opponent, thereby causing them to change their behaviour so that it aligns with the coercer's political, social, and economic preferences.[42] Coercion too, as that which deterrence seeks to oppose, is a necessarily political act. Indeed, the deterrence/coercion distinction is definitionally semantic: coercion amounts to using/threatening force to deter an opponent from continuing their present course of action, just as deterrence consists of using/threatening force to counter-coerce an opponent away from their intended coercion.[43] Cyberattack, meanwhile, is a form of coercion – at least where it seeks to change behaviour, as discussed above.

The implication is that cyber-coercion is necessarily defined by the promotion of (intrinsically political) interests. And it is this that solves cyberattack's interest-attribution problem, even where specific attacker identity remains obscured. For in order to advance a set of political interests via coercion, a coercer must necessarily identify what those interests are. To coerce without revealing desired behavioural change is contradictory. Yet in identifying the desired change in behaviour, a coercer also necessarily identifies some set of interests that it values. Wholly concealing such interests would involve concealing the desired change in behaviour, thus failing to produce coercive effect. Attempted coercion thereby serves as a preference revelation mechanism, whereby an attacker's interests are exposed by the preferred change in behaviour that is sought. The specific meaning of attempted coercive signals can be ambiguous, misinterpreted, unintentionally (de-)escalatory, or even missed altogether, moreover, which is one reason why cyber-coercion will remain difficult to use successfully.[44] But in cases where an aspiring coercer has escalated to unmissable cost-imposition – that is, coercion has actually taken place, in terms of generated effect – then a set of associated interests must necessarily be identifi*able* (even if not easily identifi*ed*, as discussed below).

Such preference revelation may make the identity of a cyber-coercer clear, even if the cyberattack itself cannot be traced via technical means.[45] If the interests that the cyber coercion seeks to advance align solely with the unique identifiable interests of a particular state/group, then that actor will have identified itself. Similarly, if a cyber aggressor exploits its successful attack via non-cyber means – say, conventional military action whilst an opponent's command/control systems have been taken temporarily offline by a cyberattack[46] – this too may reveal the attacker's identity. Yet even where a specific attacker is *not* identifiable via these revealed preferences, such preference-revelation is *still* useful for deterrence: while there may be a dozen potential culprits behind a given cyberattack, the very thing that makes them hard to distinguish – their similar interests – is *also* the thing that unifies them. Thus, while the precise identity of an attacker may remain unclear, this does not preclude an effective countervalue deterrent posture. The interests the coercion sought to advance can necessarily be identified as something the attacker – whoever it was – values, and held at risk via the threat of retaliation.[47]

The single most high-profile cyberattack to date represents a key example of interest-attribution in the absence of specific-attacker identification. Iran did not need state-of-the-art cyber forensics to have a good idea that the Stuxnet attack on its nuclear facilities uncovered in 2010 was launched by Israel, the United States, or conceivably one of those two states'

allies. While more than one technically-capable actor had an interest in retarding Iran's nuclear program, making it impossible to identify the *specific* attacker based on desired behavioural change alone, there were *interests* common to all possible attackers – the security of the United States' Middle Eastern allies – that Iran could have held at risk to achieve deterrence. Stuxnet was not an attack of explicit coercive signalling, of course; it was covert attrition, intended to degrade a capability. Nonetheless, as per the argument above, such attrition was then *exploited* coercively in non-cyber domains – and once discovered (as it was always likely to be), the *interests* of its progenitor(s) were not hard to discern (even as specific attacker *identity* remained opaque). Subsequent interest-punishment was indeed attempted by Tehran, moreover, as discussed below; the principal barrier to Iran achieving deterrence – as with Syria's inability to deter reported cyberattack on its air defences in 2007 – was not the anonymity of the cyberattack, but its military inferiority and associated paucity of sufficiently credible retaliatory options. More generally, we may struggle to identify whether the cyberattacks on Estonia (2007), Georgia (2008), Finland (2013), and Ukraine (2014) were conducted by the Russian state itself or pro-Russian 'non-state' actors with curiously state-like capabilities. Regardless of the precise attributability of the Kremlin, however, the interests at stake were clear. Deterrence failed for various reasons in these cases, but a lack of punishable revealed preferences was not one of them.

While such cases are illustrative, however, they also expose the limits of inductive case-study analysis and the value of deductive inference on this topic, because the cases' very observability is necessary evidence of deterrence failure – like much deterrence analysis, the subject suffers intrinsic selection bias.[48] The unknown number of cyberattacks contemplated-but-not-conducted due to the likely costs of potential retaliation are actually the most important measure of whether deterring technically-untraceable cyberattacks via the threat of retaliation is possible. Yet such cases necessarily remain unobservable.

Caveats and Qualifications

This argument suggests grounds for optimism about states' and their citizens' prospects for deterring cyberattacks via both cyber and non-cyber retaliatory means – at least in countries powerful enough to threaten meaningful costs. Five broad categories of caveats must be added, however, which counsel against wholesale reliance on countervalue deterrent options.

First, cyberattacks conducted wholly for the enjoyment of disruption, without intention to change behaviour – attacks by individuals/groups who derive utility from hacking itself, rather than from its potential coercive payoff – are unlikely to be deterred through the

mechanism described above. Since such individuals are not seeking to advance a political purpose, they have no identifying cause – and may thus be genuinely anonymous, if their activities cannot be technically traced, making them impossible to retaliate against. The same may be true of those who hack solely to bring private information to public attention (although even this is a form of political interest that could, hypothetically, be held at risk).[49] And the same may also be true of those who hack solely for financial enrichment, absent any higher coercive purpose. Such attacks fall beyond the purview of the deterrent posture described here – although, fortunately, such hackers are less likely to possess the wherewithal to carry-out the mass-destructive/mass-casualty cyberattacks that states or well-funded non-state groups could attempt.[50]

Second, deterrence by denial (hardening ICT systems) may have superior humanitarian consequences to deterrence by punishment, and thus suffer fewer credibility gaps. That punishment is feasible does not preclude denial being the first-choice option. For deterrence by countervalue punishment relies on having both the capability and resolve to harm that which an attempted coercer values. And such interests are likely to include civilian populations and their economies/societies – as well as various less 'worthy' causes – the targeting of which may be morally questionable.

Such concerns are not 'just' an ethical problem. Because of the moral and associated international public-opinion considerations involved, a potential cyber aggressor might assess that their target would be unwilling to impose the humanitarian costs associated with countervalue retaliation.[51] The aspiring deterrer's retaliatory threats may thus not be credible, and potential aggressors undeterred. Deterrence by denial is therefore still likely to be a preferable first-choice: it avoids the potential humanitarian consequences of a countervalue retaliatory strike, and correspondingly avoids the risk of deterrence failure through non-credible threats.[52] That said, deterrence by denial can *also* fail if perceived as non-credible, in which case the threat of punishment may achieve what the threat of denial cannot.[53] The point is simply that deterrence via threat of retaliatory punishment should be viewed as complementary to denial, rather than a silver-bullet alternative.

Though deterrence via the threat of countervalue retaliatory punishment can be ethically problematic, furthermore, all contemporary major powers' strategic postures still ultimately rest on it (i.e. the threat of nuclear retaliation). The non-credibility of deterrent threats through humanitarian 'self-deterrence' is avoided by ensuring a linkage between lower levels of violence and an escalatory spiral to higher levels of violence: lower-level aggression is deterred by the *risk* of escalation to unacceptable costs. As such, while the

humanitarian consequences of retaliation against a cyberattack would indeed represent disproportionate overkill if an attacked major power used all its retaliatory wherewithal straight away, such victims can instead choose initially to only use a proportionate level of retaliation – provided that they have an adequate spectrum of retaliatory options, obviously – thereby reducing both humanitarian and credibility problems (especially if initially restrained to counterforce actions). Subsequent counter-retaliation may escalate levels of damage, but the threat of such escalation would have a powerful chilling effect against ever reaching that position – and if conflict did escalate nonetheless, the credibility problem would dwindle anyway, since the balance of retaliatory resolve tends to lie with the side with the most to lose.[54]

Relatedly, much malicious cyber activity takes place at the level of 'irritating nuisance' rather than 'act-of-war',[55] so aspiring deterrers will have to trade-off their desire to deter major attacks against their desire to avoid potentially dangerous escalation over minor infringements.[56] Aspiring deterrers thus have to decide the level at which to set their retaliation threshold, how to gradate it, and whether to be explicit about its location. A low, ambiguous threshold will produce more deterrence, but higher risk of unwanted escalation; a high, explicit threshold will reduce the risk of potentially catastrophic escalation, but could give free-rein to nuisance attacks below the threshold.[57] While this is an important practical consideration for the real-world implementation of any deterrent posture, however, it does not undermine the overall conceptual point. Moreover, it is a problem that can be reduced via the possession of a broad spectrum of potential retaliatory options, allowing an appropriately measured response.

Third, a particularly effective cyberattack could undermine its victim's cross-domain capability to retaliate against cyberattack, which would pose a critical challenge to deterrence based on the threat of *post hoc* retaliation.[58] In a nightmare scenario for an aspiring deterrer, a devastating cyberattack might compromise nuclear command/control, removing the 'top rung' of the escalatory deterrent ladder on which the credibility of the rest depends.[59] Lower down, a cyberattack on the internet-connected logistical systems of an opponent's forces might compromise that opponent's ability to sustain a large-scale conventional military campaign.[60] Deterrence might then break down, because the only options for the aspiring deterrer may be (compromised) cyber/conventional operations – which might not be sufficient to convince an aggressor that the costs of their attack would exceed the benefits – *or* escalation straight to nuclear retaliation, which might be so extreme as to lack credibility. And in still another variant of this concern, cyber-enabled subversion/disinformation that

spread domestic discontent and division in the targeted state might undermine the strategic consensus underpinning deterrent posture and thereby remove the resolve necessary for deterrent credibility. In short, an attack that reliably undermined its target's ability to retaliate – and in which the potential aggressor knew the effectiveness of its attack in advance – could not be deterred via the threat of retaliatory punishment.

Fortunately, such attacks – including high *ex ante* certainty over effectiveness – would be hard to achieve.[61] Still, with each side knowing the other's incentives to compromise its command/control systems, and to counter-move on the assumption of such vulnerability in themselves and others, this could become a perilous source of crisis instability, with escalation taking on its own dangerously self-reinforcing dynamics.[62] Indeed, such concerns highlight an important axis along which the two post-2018 US cyber posture innovations outlined earlier must be demarcated. The 'Cyber Deterrence Initiative' – Washington's ambition to coordinate retaliation against cyberattack among its allies, thereby strengthening the effectiveness of retaliatory action – bolsters second-strike credibility. It may suffer all of the usual alliance challenges of free-riding and collective action failure, therefore – as well as offering no panacea to cyber-retaliation's potential 'collateral damage' problem – but should at least do little to increase Western adversaries' first-strike pressures. 'Persistent Engagement', however – Washington's stated intention to offensively pre-empt possible hostile cyber-action – could carry quite different implications for crisis instability, potentially incentivising US adversaries to strike first rather than risk waiting to be struck first themselves (and forcibly disarmed in the process). This does not mean that there can never be a sound defensive basis for offensive cyber-action, of course; nonetheless, those developing such postures should reflect on the ambiguous net security consequences of the counter-incentives that they create.

Variations of the 'undeterrability' caveat apply to attempts to deter two other types of cyber 'attack': those occurring once conflict is already underway, and those intended solely to gather intelligence. The latter are not strictly cyber*attacks*, since their very purpose is to go undiscovered while secretly harvesting data, although they are part of the broader cyber dimension of contemporary interstate competition. Key examples include the 'Titan Rain' and 'GhostNet' infiltrations attributed to China (although these examples can only be observed because they were discovered, and therefore are no longer effective – any active cases necessarily remain undiscovered).[63] Cyber data theft for private profit – by both criminal groups/individuals, and by governments (or state-sanctioned privateers) to enhance the competitive position of favoured corporations (such as in the 2014 U.S. lawsuit against

China)[64] and/or the overall national economy – may also fall in this category. Something reliant on going undiscovered cannot be used as a coercive lever. However, the information gathered could be valuable to some other method of coercion, thus the distinction between espionage and coercion in cyberspace is not clear-cut.

Consider cyber-theft of U.S. F-35 aircraft technical data incorporated into China's own aircraft and radar designs – systems subsequently deployed to counter U.S. forces, thereby revealing a set of deterrable coercive preferences.[65] As long as cyber-spying goes unrevealed as coercive behaviour, its progenitors may not be identifiable (and subsequently deterrable). But once the information is used, the coercive purpose is exposed.[66] Such espionage-enabled coercion may not always be as readily observable as in the high-profile F-35 case. Data theft from the likes of General Dynamics or Rolls-Royce could enable Beijing to build quieter and thus more survivable nuclear submarines, say, thereby expanding Chinese coercive options – a more secure nuclear retaliatory arsenal would expand Beijing's scope for confrontational behaviour in other domains/theatres – but without an 'obvious' F-35-esque Western copy revealing that cyber-spying had taken place. Nonetheless, a sudden step-change in Chinese submarine reactor silencing, especially one that betrayed Western design characteristics, would indicate what had happened – yet if Beijing chose to *forego* such a step-change, to *conceal* its data-theft, it would also forego the very advantage it had sought. If the targeted company detected a system breach, moreover – and few hackers could be wholly confident of avoiding *some* trace of detection – then the combination of detected-breach-plus-capability-improvement would be even clearer.

The former, meanwhile – cyberattacks conducted as part of open hostilities – may be uninterested in concealing their provenance; their purpose is to degrade opposition capabilities in pursuit of victory.[67] Such attacks do not suffer the anonymity problem, since the aggressor's identity is obvious.[68] Yet since kinetic conflict is already underway, there has already necessarily been at least a partial deterrence failure. Both sides may still refrain from escalatory cyberattacks through desire to keep conflict limited.[69] But there is also a chance that they will not, since they are already engaged in hostilities and experiencing 'retaliation' anyway.

Fourth, relying on deterrence that identifies targets for retaliation via technically-anonymous aggressors' coercive goals risks devious third-parties tricking their enemies into mutual conflict by 'framing' them: so-called 'false-flag' operations.[70] If Russia decided that its strategic interests would be served by China and the United States weakening each other militarily, for example, it might conduct a technically-anonymous cyberattack on U.S. forces

in Asia that seemingly benefitted China – or even one with ostensibly 'Chinese' technical characteristics – leading Washington to believe Beijing was opening hostilities against its regional interests. This could provoke swift U.S. retaliation, and equally swift Chinese escalatory re-retaliation, through each side's fear of losing first-move advantages.[71] Moreover, even without successful 'framing' of an innocent party as the target for misdirected retaliation, the utilisation of 'false-flags' may sew doubt and subsequently preclude effective retaliation, thereby undermining the credibility of punishment threats.

This problem is more context-specific than insurmountable, however. Between two parties with a less offence-dominant strategic relationship than the contemporary U.S.-Chinese situation is often taken to be[72] – and even that is contestable on the cyber front[73] – a lack of first-move advantages could allow scope for diplomatic consultation over the origin of the attack.[74] If the 'framed' aggressor repudiated the coercive goals that appeared to point in its direction – if it consciously eschewed the potential benefits of the apparent coercion – that would serve as a costly signal that it was not to blame.[75] The third-party attempting such framing would then itself be in a perilous situation: if its own subversive agenda is identified as the next-most-likely payoff of the attack, the two parties it had attempted to trick into conflict could retaliate in concert. This is also – as with anonymity/attribution – not a problem unique to cyber, and has been surmounted in other strategic domains.[76] As such, while a potential problem for deterrence, the threat of third-party 'framing' is not a risk-free option for a potential 'framer', and can be minimized if potential adversaries pursue deterrent postures that minimize first-move advantages (as has already been achieved at the 'top rung' of the escalatory ladder, via the major powers' survivable nuclear arsenals, albeit less so at the lower-level intersection of cyber and conventional conflict).

A similar critique and response applies to the possibility of 'normal' cyber accidents – things simply going wrong with ever-more-complex ICT systems, without any intentional malicious action, thanks to inadvertent failures in their design, manufacture, or upkeep – being falsely construed as attacks.[77] This poses a challenge to deterrence, since it might lead to either insufficient retaliation (if an attack was mistaken for an accident) or mis-retaliation (if an accident was mistaken for an attack). Again, however, identifying the coercive interests at stake can help to mitigate the accident-or-attack problem. The repudiation of the coercive benefits of an accident could serve as a costly signal of blamelessness. A lack of repudiation of such gains, by contrast, would signal that it should be treated as a coercive attack and the associated interests held at risk, even if the attacker's identity remains unclear.

Finally, following from the caveat on false-flags and accidents, recognizing that cyber coercion will always seek to advance some set of interests does not mean that discerning such interests will be *easy*. On the contrary, for the same reasons that aggressors may avail themselves of the anonymity potential of cyberspace – to reap the benefits without bearing the retaliatory consequences – they may seek to conceal the interests being advanced. In the context of Stuxnet, for example, had that attack actually emanated from Saudi Arabia (say), it would not be wrong to say that Iran had been attacked by 'America's Middle Eastern allies' (to use this article's earlier phraseology). If Tehran chose to misguidedly punish Israel *specifically*, however, that might be a perfectly satisfactory outcome for Riyadh. In short, there could be more than one *set* of inter*ests* – united by *a* certain inter*est* (opposition to an Iranian nuclear capability), but not identical in other respects – being advanced by the same attack.

This, then, is another area where the ideal-type distinction between coercion and espionage in cyberspace becomes hard to sustain. Coercers will often *want* to be correctly attributed, to maximize the clarity of intended signals.[78] But many potential cyberattackers will prefer to achieve their strategic goals via 'cyber covert action' rather than overt cyber coercion – just as states have often opted for 'conventional' covert action in their quest for deniability – to gain benefits without possible retaliatory costs.[79] They would still be pursuing some strategic ends via cyber coercion, but hoping to hide their true purpose among many diffuse, overlapping, seemingly contradictory possible outcomes of a given cyberattack, such that the target of the attack would find it hard to identify the interests actually being advanced among all the 'noise'. Accordingly, those seeking to hold the relevant interests at risk to generate deterrence may have to look at seemingly indirect consequences, apparently removed from the initial cyberattack, to identify the set of interests being advanced.

Crucially, however, whereas the identity of an attacker may be erased such that even state-of-the art cyber forensic work cannot discern it, the *interests* being advanced can only ever be obscured (just as with 'conventional' covert action), otherwise they would have to not be advanced at all. Moreover, with enough relative power and resolve, there is no reason in principle why the target of such an attack could not hold *all* of the interests seemingly being advanced at risk – and the more severe the suffered attack, the easier it is to make expansive retaliatory threats credible[80] – removing the problem of having to choose which was the 'real' set.[81] Indeed, to the extent that Iran retaliated against Israel (via its Hezbullah and Hamas proxies), America (via the 2012 US financial hack), *and* Saudi Arabia (via the 2012 Saudi Aramco cyberattack), this appears to be the route that Tehran chose following Stuxnet. Such

an approach can be counterproductive, of course, particularly if/when the interests advanced were seemingly contradictory and given the potential downsides to retaliating against the wrong opponent (opprobrium-inducing injustice and enmity-creating blowback). Nonetheless, this is not a problem in principle with the notion of retaliating against all identifiable interests; it is simply a contingent and case-specific possibility to be weighed in a given scenario, with ambiguous net ramifications. Lindsay, for example, notes that punishing many for the crimes of a few can be 'unpopular', 'illegitimate', and 'counterproductively embolden[ing]' of opponents.[82] This does not, however, mean it *must* yield a net negative strategic outcome: with high enough stakes and enough relative power, illegitimacy, unpopularity, and the creation of new enemies may be prices worth paying to ensure the underlying culprits suffer punishing costs as part of ensuing retaliation – as per parallels in counterterrorism and counterinsurgency.[83]

More practically, therefore, while many sets of interests may seemingly be advanced as part of the attacker's attempts to conceal their true purpose, the sets of interests being advanced the *most* by an attack should be loosely rank-orderable[84] – or at least, not obviously *not* so to a potential aggressor weighing costs/benefits *ex ante*. Furthermore, if the potential beneficiaries of the interests ranked as being advanced the most by an attack refused to repudiate the coercive gains, this would be a costly signal that those were the interests being advanced and that should therefore be retaliated against. If such gains were repudiated, by contrast, that would be a costly signal that the previously first-most-likely-seeming set of interests were not the ones being advanced, and the party seeking deterrence could progress to holding the second-most-likely set of interests at risk – and so forth. Thus, while the coercive interests at stake may not always be *simple* to identify, this does not mean that they are not always identifi*able* – a significant improvement over attacker identity, which may indeed not be identifiable at all.

Conclusion

The technical anonymity of cyberattack is commonly cited as the principal barrier to the deterrence of such attacks: absent a 'return address', how could a victim know where to target its retaliation? Consequently, approaches to security against such attacks have focused on hardening ICT networks: that is, deterrence/defeat by denial. Such hardening and resilience-building remains worthwhile, moreover, because the threat of retaliation against cyberattack is not a deterrent panacea; problems of humanitarian consequences, military/technical vulnerability, and possible deception all challenge the credibility of retaliatory threats. More

resilient systems should also strengthen crisis stability by lowering the escalatory pressures associated with decisive first-move advantage.

Nonetheless, the technical anonymity of cyberattack is not the fundamental obstacle to deterrence via the threat of punishment that it is often supposed to be, for attacker *identity* and attacker *interests* – while often conflated – are not the same thing. On the contrary, because cyberattack is predominantly a coercive act intended to advance political goals, the aggressor's interests can be identified via the attempted coercion – certain caveats notwithstanding – and thus held at risk by an aspiring deterrer with sufficient resolve and capability. Attempted coercion is itself a preference-revelation mechanism, in short, and those revealed preferences can be retaliated against even when the coercer themselves escapes attribution.

This approach is not without peril; retaliating against 'innocent' states and/or civilians carries normative costs and blowback risks, while cross-domain escalation may become dangerously unstable, especially if policymakers acquire false confidence in the efficacy of offensive cyber operations. Nonetheless, as the damage and lethality potential of future cyberattack rises in line with the sophistication of cyber weapons, governments and the populations they seek to protect may draw comfort from the fact that deterrence via the threat of retaliatory punishment – which, for all of its ethical questions, remains the final backstop of major powers' strategic postures – holds some of the same promise against cyber threats that it has against the nuclear and conventional threats of the present and recent past.

References

**All URLs accessible as of 23 December 2019*

[1] On cyber capability growth, see: Dale Peterson, 'Offensive Cyber Weapons: Construction, Development, and Employment', *Journal of Strategic Studies* 36:1, 2013, pp. 120-24.

[2] Some have derided discussions of cyber deterrence as misguidedly applying an outdated 'Cold War paradigm' to an ill-suited strategic context, e.g. Daniel Steed, 'Cyber war, let's get real(ist)', *WarOnTheRocks.com*, 14 October 2013, https://warontherocks.com/2013/10/cyber-war-lets-get-realist/. There is nothing 'Cold War' about deterrence, however, despite the nuclear era bringing the concept newfound prominence. Rather, it is a timeless and necessary attribute of interactions between mutually armed parties under anarchy: Ben Buchanan, 'Cyber Deterrence Isn't MAD; It's Mosaic', *Georgetown Journal of International Affairs: Cyber IV*, 2014, p. 131. It is therefore an entirely appropriate concept for the cyber age.

[3] Sorin Dumitru Ducaru, 'The Cyber Dimension of Modern Hybrid Warfare and Its Relevance for NATO', *Europolity* 10:1, 2016, pp. 7-23.

[4] Richard L. Kugler, 'Deterrence of Cyber Attacks', in Franklin D. Kramer, Stuart H. Starr, and Larry K. Wentz (eds.), *Cyberpower and National Security* (Dulles: Potomac Books, 2009), pp. 309-41; Martin C. Libicki, *Cyberdeterrence and Cyberwar* (Santa Monica: RAND, 2009), p. 44; Thomas Rid and Ben Buchanan, 'Attributing Cyber Attacks', *Journal of Strategic Studies* 38:1-2, 2015, pp. 4-37.

[5] Erica D. Borghard and Shawn M. Lonergan, 'The Logic of Coercion in Cyberspace', *Security Studies* 26:3, 2017, pp. 452-81.

[6] Key recent works on cyber deterrence/coercion share this perspective. Jon R. Lindsay argues that 'cyber operations are unsuited for coercive signaling': 'Tipping the Scales: The Attribution Problem and the Feasibility of Deterrence against Cyberattack', *Journal of Cybersecurity* 1:1, 2015, p. 54. Borghard and Lonergan similarly contend that 'Signaling in cyberspace is the most problematic of all the domains': 'The Logic of Coercion in Cyberspace', p. 456.

[7] Robert A. Pape, *Bombing to Win: Air Power and Coercion in War* (Ithaca: Cornell UP, 1996), p. 4.

[8] As Borghard and Lonergan's own Schelling-derived caveat admits, coercion need not take the form of explicit threat-issuance to nonetheless be operative; key is simply that it is *inferred* by targets, based on their assessment of others' capabilities, interests, and behaviours: 'The Logic of Coercion in Cyberspace', p. 455 (n.11). Indeed, cyber capabilities may yield coercive options well beyond mere signalling/threat-issuance: David Betz, 'Cyberpower in Strategic Affairs: Neither Unthinkable nor Blessed', *Journal of Strategic Studies* 35:3, 2012, pp. 689-711.

[9] Lindsay, 'Tipping the Scales', p. 59. This is not a denial that states are using cyber exploits against each other all the time, but simply a recognition that no major power has yet (as of December 2019) suffered a cyber-induced mass casualty attack (2017's Pyongyang-backed WannaCry attack perhaps comes closest to date, insofar as it disrupted various national healthcare systems and thereby harmed a 'mass casualty' number of patients).

[10] Albeit utilizing embedded induction from prior patterns of deterrence behaviour: David Blagden, 'Induction and Deduction in International Relations: Squaring the Circle between Theory and Evidence', *International Studies Review* 18:2, 2016, pp. 195-213.

[11] Christopher H. Achen and Duncan Snidal, 'Rational Deterrence Theory and Comparative Case Studies', *World Politics* 41:2, 1989, pp. 143–69; Lindsay, 'Tipping the Scales', p. 54.

[12] Thomas Rid and Peter McBurney, 'Cyber-Weapons', *RUSI Journal* 157:1, 2012, p. 7.

[13] Buchanan, 'Cyber Deterrence Isn't MAD', p. 131. Note too that cross-/multi-domain deterrent and/or defensive strategies, i.e. those seeking to protect more than solely cyber interests (and utilising more than solely cyber capabilities), might themselves contain an *offensive* cyber element: Michael Fischerkeller, 'Incorporating Offensive Cyber Operations into Conventional Deterrence Strategies', *Survival* 59:1, 2017, pp. 103-134. Such offensive cyber actions against non-cyber attacks are not the article's focus, however. Moreover, misplaced faith in the utility of offensive cyber operations could itself become a dangerous source of unneeded escalatory pressures, as discussed subsequently: Brandon Valeriano and Benjamin Jensen, *The Myth of the Cyber Offense: The Case for Restraint* (Washington: Cato Institute, 2019), https://www.cato.org/sites/cato.org/files/pubs/pdf/pa862.pdf.

[14] Robert Jervis, *The Meaning of the Nuclear Revolution: Statecraft and the Prospect of Armageddon* (Ithaca: Cornell UP, 1989), pp. 8-14; Pape, *Bombing to Win*, pp. 13-14.

[15] The risk of global commerce/communications disruption may also 'deter' potential cyberattackers, if they fear the self-harm of such disruption: Joseph S. Nye, 'Deterrence and Dissuasion in Cyberspace', *International Security* 41:3, 2016-17, pp. 58-60. However, this article's focus is attacks that aggressors would have otherwise judged to be advantageous *absent* their opponent's deterrent strategies, all else held equal, rather than those foregone because of other concerns.

[16] One caveat would be if cyberattacks became capable of compromising nuclear command/control, thus removing the risk of cross-domain retaliation, as discussed subsequently.

[17] Cited in Paul Cornish et al., *On Cyber Warfare* (London: Chatham House, 2010), p. 29.

[18] Likewise for the concern that cyber deterrence may be fragile if retaliation cannot be repeated (because of single-use/temporally-degenerative capabilities): Libicki, *Cyberdeterrence and Cyberwar*, pp. 56-59.

[19] Early strategists' understanding of the term was that the party possessing escalation dominance should be able to *prevail* militarily at each level of escalation: Herman Kahn, *On Escalation: Metaphors and Scenarios* (New Brunswick: Transaction, 2010 [1965]), pp. 23,136-37. With such 'victory' a doomed hope following the advent of nuclear mutually-assured destruction, however, the term should be properly understood as 'escalation *equivalence*': aspiring deterrers possessing sufficient capability-plus-resolve to escalate to levels of retaliation whereby potential aggressors' expected benefits never exceed their expected costs, creating deterrence through threatened escalation to ever-greater-yet-still-not-advantageous pain: Charles L. Glaser, *Analyzing Strategic Nuclear Policy* (Princeton: Princeton UP, 1990), pp. 50,55-57; Thomas C. Schelling, *Arms and Influence* (New Haven: Yale UP, 2008 [1966]), p. 104.

[20] Barry R. Posen, *Inadvertent Escalation: Conventional War and Nuclear Risks* (Ithaca: Cornell UP, 1991).

[21] Jervis, *The Meaning of the Nuclear Revolution*, pp. 35-38.

[22] Schelling, *Arms and Influence*, p. 47.

[23] Kadhim Shubber, 'A simple guide to GCHQ's surveillance programme Tempora', *Wired*, 24 June 2013, https://www.wired.co.uk/article/gchq-tempora-101.

[24] US Government, *Strategy for Operating in Cyberspace* (Washington: US Department of Defense, 2011), http://www.defense.gov/news/d20110714cyber.pdf; HM Government, *The UK Cyber Security Strategy:*

Protecting and Promoting the UK in a Digital World (London: UK Cabinet Office, 2011), https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/60961/uk-cyber-security-strategy-final.pdf.

[25] US Government, *Cyber Strategy* (Washington: US Department of Defense, 2015), http://www.dtic.mil/doctrine/doctrine/other/dod_cyber_2015.pdf, 11; HM Government, *National Cyber Security Strategy 2016-2021* (London: UK Cabinet Office, 2016), https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/567242/national_cyber_security_strategy_2016.pdf, 41-52.

[26] David E. Sanger and Elisabeth Bumiller, 'Pentagon to consider cyberattacks acts of war', *New York Times*, 31 May 2011, http://www.nytimes.com/2011/06/01/us/politics/01cyber.html?_r=1; Finbarr Bermingham, 'NATO Summit 2014: cyber-attack could trigger retaliation from all NATO members', *International Business Times*, 5 September 2014, http://www.ibtimes.co.uk/nato-summit-2014-cyber-attack-could-trigger-retaliation-all-nato-members-1464230; Anna Mikhailova, 'UK could retaliate against cyber attacks with missiles, Attorney General says', *The Telegraph*, 23 May 2018, https://www.telegraph.co.uk/politics/2018/05/23/uk-has-legal-right-retaliate-against-cyber-attacks-missiles/. The US Government's invocation of cyber-theft as a motive for its trade sanctions on China is consistent with this pattern: BBC News, 'A quick guide to the US-China trade war', 16 December 2019, https://www.bbc.co.uk/news/business-45899310.

[27] US Government, *National Cyber Strategy of the United States of America* (Washington: The White House, 2018), https://www.whitehouse.gov/wp-content/uploads/2018/09/National-Cyber-Strategy.pdf; Theresa Hitchens, 'US urges 'like-minded' countries to collaborate on cyber deterrence', *Breaking Defense*, 24 April 2019, https://breakingdefense.com/2019/04/us-urging-likeminded-countries-to-collaborate-on-cyber-deterrence/; Greg Myre, ''Persistent Engagement': the phrase driving a more assertive U.S. spy agency', *NPR*, 26 August 2019, https://breakingdefense.com/2019/04/us-urging-likeminded-countries-to-collaborate-on-cyber-deterrence/.

[28] Valeriano and Jensen, *The Myth of the Cyber Offense*.

[29] Cornish et al., *On Cyber Warfare*, pp. 42-45; see also Daryl G. Press, *Calculating Credibility: How Leaders Assess Military Threats* (Ithaca: Cornell UP, 2005), pp. 2-3.

[30] Nye, 'Deterrence and Dissuasion in Cyberspace', pp. 49-52.

[31] Certainly, progress is being made on this, particularly identifying technical 'fingerprints' of major states' cyberattacks: Stewart Baker, 'The Attribution Revolution', *ForeignPolicy.com*, 17 June 2013, https://foreignpolicy.com/2013/06/17/the-attribution-revolution/; Brian M. Mazanec and Bradley A. Thayer, *Deterring Cyber Warfare: Bolstering Strategic Stability in Cyberspace* (New York: Palgrave, 2015), pp. 57-63. Nonetheless, this is unlikely to become perfect, particularly as attackers adapt to defenders' progress, and given the potential for 'false-flag' attacks discussed below. All other fields of military-technological development have witnessed iterative back-and-forths between offensive penetration and defensive shielding, so recent innovations in cyber 'fingerprinting' will be countered by anonymization innovations.

[32] Kugler, 'Deterrence of Cyberattacks', pp. 309-41; Cornish et al., *On Cyber Warfare*; Kenneth Geers, 'The Challenge of Cyber Attack Deterrence', *Computer Law and Security Review* 26:3, 2010, pp. 298-302; Mazanec and Thayer, *Deterring Cyber Warfare*, pp. 29-43.

[33] Libicki, *Cyberdeterrence and Cyberwarfare*, pp. 41-52; Cornish et al., *On Cyber Warfare*, p. 13.

[34] Evidence suggests the Russian and Chinese governments, among others, see the use of arms-length 'non-state' cyber proxies as deniable tools of coercive statecraft: Alexander Klimberg, 'Mobilising Cyber Power', *Survival* 53:1, 2011, pp. 41-60.

[35] Daniel Moore and Thomas Rid, 'Cryptopolitik and the Darknet', *Survival* 58:1, 2016, pp. 7-38.

[36] Stuxnet provides the quintessential example: James P. Farwell and Rafal Rohozinski, 'Stuxnet and the Future of Cyber War', *Survival* 53:1, 2011, pp. 23-40.

[37] Rid and Buchanan, 'Attributing Cyber Attacks', pp. 30-31.

[38] Borghard and Lonergan, 'The Logic of Coercion in Cyberspace', p. 459; Libicki, *Cyberdeterrence and Cyberwar*, p. 128.

[39] Glenn H. Snyder, *Deterrence and Defense: Toward a Theory of National Security* (Princeton: Princeton UP, 2015 [1961]), pp. 14-16.

[40] Carl von Clausewitz, *On War* (Princeton: Princeton UP, 1976 [1832]), p. 87.

[41] Cornish et al., *On Cyber Warfare*, p. 12.

[42] Pape, *Bombing to Win*, p. 12. 'Compellence' is sometimes treated separately from coercion; since it essentially represents 'total' coercion, however – the target has no choice but to acquiesce – it is subsumed here within coercion: *Ibid.*, p. 4; Schelling, *Arms and Influence*, pp. 69-91.

[43] Lawrence Freedman, 'Deterrence: A Reply', *Journal of Strategic Studies* 28:5, 2005, pp. 789-790.

[44] Borghard and Lonergan, 'The Logic of Coercion in Cyberspace', pp. 455-459.

[45] Kugler, 'Deterrence of Cyberattacks', pp. 309-310.

[46] Erik Gartzke, 'The Myth of Cyberwar: Bringing War in Cyberspace Back Down to Earth', *International Security* 38:2, 2013, p. 43.

[47] Holding interests at risk *without* specifically-identified attribution goes beyond valuable recent work on interests as a *route* to attribution, e.g. Rid and Buchanan, 'Attributing Cyber Attacks', p. 8.

[48] Lindsay, 'Tipping the Scales', p. 54.

[49] Both 'recreational' and 'hacktivist' attackers *do* have interests that can be jeopardized, of course, but doing so via punishment is challenging. The former values the opportunity to commit a successful attack for the mere thrill of it (making punishment and denial one-and-the-same). The latter values something that the wider population may value too (i.e. transparency in public life), so punishing their interests could involve punishing the people that the state exists to protect.

[50] Certain capability gaps between individual hackers and state-level cyber forces have indeed closed over time; consider reduced need for state-level supercomputers to break encryption, due to bot technology: Libicki, *Cyberdeterrence and Cyberwar*, pp. 47-48. Nonetheless, as the technological sophistication frontier is pushed out by those with the greatest resources (powerful states and their corporations foremost among them), the chances of sparsely-resourced individuals/groups – however talented – keeping pace becomes small. Such state-versus-individual capability disparities are thus likely to remain on aggregate, even if gifted hackers close gaps in certain areas. Stuxnet, in particular, shows the resources required – including a replica of the entire targeted infrastructure, and an expert human intelligence operation to penetrate air-gapped systems – to propagate an attack of even modest state-level strategic significance: Jon R. Lindsay, 'Stuxnet and the Limits of Cyber Warfare', *Security Studies* 22:3, 2013, pp. 365-404.

[51] Sarah Kreps and Jacquelyn Schneider, 'Escalation Firebreaks in the Cyber, Conventional, and Nuclear Domains: Moving Beyond Effects-Based Logics', *Journal of Cybersecurity* 5:1, 2019, pp. 1-11. Libicki sees the risk of third-party censure/opprobrium as a fundamental challenge to the credibility of deterrence-by-retaliation in response to cyberattack in a way that it has not been for the deterrence of nuclear attack, particularly on the grounds that numerous capable third-parties may join the fight: Libicki, *Cyberdeterrence and Cyberwar*, p. 42; see also Lindsay, 'Tipping the Scales', p. 57. Yet while facing normative opprobrium may indeed impact a state's retaliatory resolve, there is no clear reason why third-parties would join hostilities against a retaliating state – opening themselves to re-retaliation in the process – simply to censure it for retaliating. A related argument, albeit with different causal underpinnings, is that an attack's victim may be self-deterred from certain retaliatory cyber options through fear that they would result in fratricide of their own ICT systems: *Ibid.*, p. 57. This is a relevant concern, but one contingent on the cyber-weapon in question – and on aspiring deterrers' weighting of absolute versus relative gains/losses – rather than a fundamental barrier to retaliation *per se*.

[52] Threatening *ex ante* success-prevention may be inherently more credible than threatening *ex post* punishment, since the former promises minimized costs on both sides: Snyder, *Deterrence and Defense*, p. 16. Deterrence by punishment also cedes control of decision – over how much pain to bear – to one's opponent, and may therefore also be less precise/reliable when dealing with aggressors of uncertain cost-tolerance/risk-acceptance: Lawrence Freedman, *Deterrence* (Cambridge: Polity, 2004), p. 39.

[53] Geers, 'The Challenge of Cyber Attack Deterrence', pp. 300-1.

[54] Jervis, *The Meaning of the Nuclear Revolution*, pp. 29-35; Schelling, *Arms and Influence*, pp. 69-91,99-105,172-76; Snyder, *Deterrence and Defense*, p. 226.

[55] Erik Gartzke and Jon R. Lindsay, 'Weaving Tangled Webs: Offense, Defense, and Deception in Cyberspace', *Security Studies* 24:2, 2015, pp. 316-48.

[56] Lawrence J. Cavaiola, David C. Gompert, and Martin Libicki, 'Cyber House Rules: On War, Retaliation and Escalation', *Survival* 57:1, 2015, pp. 81-104.

[57] Lindsay, 'Tipping the Scales', pp. 58-64. This trade-off exists throughout deterrent posture – 'asymmetric escalation' doctrines deliver much deterrence but with escalatory dangers, for example, while 'assured retaliation' doctrines offer less deterrence of low-level irritations but also fewer escalatory risks: Vipin Narang, *Nuclear Strategy in the Modern Era: Regional Powers and International Conflict* (Princeton: Princeton UP, 2014), pp. 17-20.

[58] James J. Wirtz, 'The Cyber Pearl Harbor', *Intelligence and National Security* 32:6, 2017, pp. 758-767. Of course, this is not to suggest that pulling off such a successful attack – or achieving the strategic shock/paralysis that it might induce – would be easy or even feasible, but it would clearly be an attractive prospect for an aspiring aggressor: James J. Wirtz, 'The Cyber Pearl Harbor Redux: Helpful Analogy or Cyber Hype?', *Intelligence and National Security* 33:5, 2018, pp. 771-773.

[59] Geers, 'The Challenge of Cyber Attack Deterrence', p. 301; Andrew Futter, '*War Games* Redux? Cyberthreats, US-Russian Strategic Stability, and New Challenges for Nuclear Security and Arms Control', *European Security* 25:2, 2016, pp. 163-80; Erik Gartzke, 'Thermonuclear Cyberwar', *Journal of Cybersecurity* 3:1, 2017, pp. 37-48; Andrew Futter, *Hacking the Bomb: Nuclear Weapons, Cyber Threats, and the Incipient Digital Age* (Washington, DC: Georgetown UP, 2018).

[60] David C. Gompert and Martin Libicki, 'Cyber Warfare and Sino-American Crisis Instability', *Survival* 56:4, 2014, pp. 10-11.

[61] A common characteristic of cyberattack is uncertainty even on the attacker's part over precisely what effects their action will produce: Borghard and Lonergan, 'The Logic of Coercion in Cyberspace', p. 456.

[62] Ben Buchanan, *The Cybersecurity Dilemma: Hacking, Trust, and Fear between Nations* (New York: Oxford UP, 2017); Herbert Lin, 'Escalation Dynamics and Conflict Termination in Cyberspace', *Strategic Studies Quarterly* 6:3, 2012, p. 57. For theoretical background, see: Posen, *Inadvertent Escalation*, pp. 2-3; Clausewitz, *On War*, pp. 75-89.

[63] Cornish et al., *On Cyber Warfare*, pp. 8-9.

[64] BBC, 'US Justice Department charges Chinese with hacking', 19 May 2014, http://www.bbc.co.uk/news/world-us-canada-27475324.

[65] Sydney J. Freedberg, 'Top official admits F-35 stealth fighter secrets stolen', *Breaking Defense*, 20 June 2013, http://breakingdefense.com/2013/06/top-official-admits-f-35-stealth-fighter-secrets-stolen/; Marcus Weisgerber, 'China's copycat jet raises questions about F-35', *Defense One*, 23 September 2015, http://www.defenseone.com/threats/2015/09/more-questions-f-35-after-new-specs-chinas-copycat/121859/; Dave Majumdar, 'Did China just make all stealth fighters (think the F-22 and F-35) obsolete?', *National Interest*, 27 February 2017, http://nationalinterest.org/blog/the-buzz/did-china-just-make-all-stealth-fighters-think-the-f-22-f-35-19608.

[66] Indeed, since cyber exploits that are revealed via attack may be curtailed as sources of future information/leverage, this creates a paradox. For those capable of both gaining covert cyber access *and* utilizing the information gathered in non-cyber domains, it may prove beneficial to allow adversaries unfettered action in cyberspace – even at the expense of regular nuisance-level cyberattacks – so as to preserve access to exploitable information. Again, this points towards cyberspace as a reinforcer of the balance of interstate power – and again, it suggests that the most strategically consequential dimensions of cyber capability are those that go unobserved: Gartzke and Lindsay, 'Weaving Tangled Webs', pp. 316-48.

[67] On the distinction between pursuing military victory and coercion, see: Pape, *Bombing to Win*, pp. 13-15.

[68] Although opportunistic third-parties not already involved in the conflict could use such hostilities as cover for their own agenda against one/multiple protagonists.

[69] Few wars escalate into wholly unbounded exertions of brute force, rather than still-bounded cost-imposition, which is why 'coercion' versus 'victory' remains another falsely-dichotomous ideal-type. Indeed, even the archetypal 'total war' still featured discretionary limits: Patrick Porter, 'A Matter of Choice: Strategy and Discretion in the Shadow of World War II', *Journal of Strategic Studies* 35:3, 2012, pp. 317-43.

[70] Libicki, *Cyberdeterrence and Cyberwar*, p. 44.

[71] Gompert and Libicki, 'Cyber Warfare and Sino-American Crisis Instability', pp. 8-10; Joshua Rovner, 'Two Kinds of Catastrophe: Nuclear Escalation and Protracted War in Asia', *Journal of Strategic Studies* 40:5, 2017, pp. 699-706.

[72] Avery Goldstein, 'First Things First: The Pressing Danger of Crisis Instability in U.S.-China Relations', *International Security* 37:4, 2013, pp. 49-89.

[73] Rebecca Slayton, 'What Is the Cyber Offense-Defense Balance? Conceptions, Causes, and Assessment', *International Security* 41:3, 2016-17, pp. 72-109; Valeriano and Jensen, *The Myth of the Cyber Offense*.

[74] Furthermore, while cyber dynamics may contribute to worrisome offense-dominance in the overall U.S.-China relationship, Western fears of relative cyber weakness vis-à-vis China may also be overblown: Jon R. Lindsay, 'The Impact of China on Cybersecurity: Friction and Fiction', *International Security* 39:3, 2014-15, pp. 7-47; Rovner, 'Two Kinds of Catastrophe', p. 712.

[75] Charles L. Glaser, *Rational Theory of International Politics: The Logic of Competition and Cooperation* (Princeton: Princeton UP, 2010), pp. 64-66; Andrew Kydd, 'Trust, Reassurance, and Cooperation', *International Organization* 54:2, 2000, pp. 325-57.

[76] One attempt at cataloguing such attacks alleges fifty-three such (admitted) attacks since the 1930s, with state use of sponsored false-flag terrorist attacks as a pretext for preferred state policy emerging as a favourite: *WashingtonsBlog.com*, '53 Admitted False Flag Attacks', 23 February 2015, http://www.washingtonsblog.com/2015/02/x-admitted-false-flag-attacks.html.

[77] Lucas Kello, 'The Meaning of the Cyber Revolution: Perils to Theory and Statecraft', *International Security* 38:2, 2013, p. 31. Such accidents become increasingly likely as system complexity increases.

[78] Borghard and Lonergan, 'The Logic of Coercion in Cyberspace', p. 459.

[79] Covert action is, 'An operation designed to influence [others] in support of foreign policy in a manner that is not necessarily attributable to the sponsoring power': U.S. Central Intelligence Agency, *Consumer's Guide to Intelligence* (Washington, DC: U.S. Central Intelligence Agency, 1995), p. 3; see also David F. Rudgers, 'The Origins of Covert Action', *Journal of Contemporary History* 35:2, 2000, pp. 249-62.

[80] Lindsay, 'Tipping the Scales', p. 63.

[81] To the extent that Iran retaliated against Israel (via its Hezbullah and Hamas proxies), America (via the 2012 U.S. financial hack), *and* Saudi Arabia (via the 2012 Saudi Aramco cyberattack), this was indeed the route that Tehran chose following Stuxnet.

[82] *Ibid.*, p. 57.

[83] In the absence of precise identification of specific insurgent/terrorist attackers, counterinsurgent/-terrorist forces may resort to punishing their manifested interests (razing an insurgent-hosting village, say, or retarding terrorists' political cause): Gil Merom, 'Strong Powers in Small Wars: The Unnoticed Foundations of Success', *Small Wars and Insurgencies* 9:2, 1998, pp. 38-63; Robert F. Trager and Dessislava P. Zagorcheva, 'Deterring Terrorism: It Can Be Done', *International Security* 30:3, 2005-6, pp. 87-123. Drawing the cyberattack/insurgency parallel, see: Lindsay, 'Tipping the Scales', p. 57. Of course, this approach can also bring heavy moral/blowback costs, but the point is that there is a rationale weighing against such costs.

[84] Where 'most' is a function of both coercive gains *and* avoided retaliatory costs.