Philos. Technol. (2018) 31:339–355
https://doi.org/10.1007/s13347-017-0290-2

RESEARCH ARTICLE

The Limits of Deterrence Theory in Cyberspace

Mariarosaria Taddeo [1,2]

Received: 17 September 2017 / Accepted: 27 September 2017 / Published online: 16 October 2017

Abstract In this article, I analyse deterrence theory and argue that its applicability to cyberspace is limited and that these limits are not trivial. They are the consequence of fundamental differences between deterrence theory and the nature of cyber conflicts and cyberspace. The goals of this analysis are to identify the limits of deterrence theory in cyberspace, clear the ground of inadequate approaches to cyber deterrence, and define the conceptual space for a domain-specific theory of cyber deterrence, still to be developed.

Keywords Cyberspace · Cyber conflicts · Defence · Deterrence · Retaliation · State · Stability · International Relations

1 Introduction

Historically, the design and deployment of new and more effective weapons (from bombs and aircraft to chemical and nuclear weapons) have often posed the need to define new strategies to deter their use. This is also the case when considering cyber weapons. Their relatively low entry cost and the high chances of success make cyber weapons an elective means for state and non-state actors to assert their authority, show their power, and prove their capabilities in cyberspace.

This poses serious risks of escalation, for the increasing use of cyber weapons invites frictions and tensions that may lead to the sparking of new cyber conflicts, which could intensify and jeopardise international stability and the security of our societies. For this reason, state and non-state actors, scholars, military strategists, and policy makers have increasingly stressed the need to develop cyber deterrence as a

This article is part of a project on "Landscaping Cyber Detterrence" funded bythe John Fell OUP Research Grant, University of Oxford (grant number 151/063)

✉ Mariarosaria Taddeo
mariarosaria.taddeo@oii.ox.ac.uk

[1] Oxford Internet Institute, University of Oxford, 1, St Giles, Oxford OX1 3JS, UK

[2] The Alan Turing Institute, 96 Euston Road, London NW1 2DB, UK

crucial step in any plan for international stability (European Union 2014; International Security Advisory Board 2014; UN Institute for Disarmament Research 2014; UK Government 2014; European Union 2015). Nonetheless, applying traditional[1] deterrence theory (henceforth simply deterrence theory) to cyberspace proves to be problematic, when not ineffective. Cyber conflicts differ radically from violent (kinetic) conflicts and define a scenario that is actually the opposite of the one for which deterrence theory was developed.

Consider Morgan's six elements of deterrence (Morgan 2003). Deterrence works in a scenario characterised by (1) a prevailing, kinetic military conflict; (2) the applicability of rational choice models to identify strategies for the involved parties; (3) positive attribution, as not problematic; (4) singular retaliation (more on this presently), as sufficient to inflict severe punishment to the opponent; (5) the possibility of a clear demonstration of the defender's capabilities; and (6) full control over retaliation. To this scenario, cyber conflicts oppose one characterised by (1) several state-run, non-kinetic cyber operations; (2) multiple (state and non-state) actors; whose (3) cost–benefit analyses vary depending on their nature; (4) non-symmetrical, multilateral interactions; (5) ever-changing dynamics; and where (6) ambiguity (rather than certainty) shapes strategies (Sterner 2011) (Haggard and Simmons 1987; Jervis 1988; Libicki 2011).

The differences between the kinetic and cyber scenario yield serious problems when applying deterrence theory in cyberspace. While there is a general consensus on what these problems are (for example, problems of attribution and proportionality), there is much less agreement on whether and how they can be solved (Kugler 2009; Tanji 2017). Some suggest that these problems are unsolvable and that the nature of cyberspace is such that deterrence will ultimately be ineffective in this domain. In this vein, Lan and colleagues stress that:

> the anonymity, the global reach, the scattered nature, and the interconnectedness of information networks greatly reduce the efficacy of cyber deterrence and can even render it completely useless (Lan et al. 2010, 1).

The opposite view holds that deterrence could play a crucial role in averting cyber conflicts and their escalation. The question is whether deterrence theory provides the right framework for cyber deterrence or a new theory of deterrence—'a new mind-set and changed expectations' (Sterner 2011, 62)—should be developed to address the specificity of cyber conflicts and cyberspace. I agree with this view and address this question in the rest of this article.

In the next sections, I will analyse the core elements of deterrence theory—attribution, defence and retaliation, and signalling—and the extent to which each of them would be effective in cyberspace. I will argue that the limits of deterrence theory in cyberspace are not trivial and indicate fundamental inconsistencies between the theory and the nature of cyber conflicts and cyberspace. The goals are to identify the limits of the application of deterrence theory to cyber conflicts (Taddeo 2016), clear the ground of inadequate approaches to cyber deterrence, and define the conceptual space for a

[1] For the purposes of this article, I will use the expression 'traditional deterrence theory' to refer to any theory of deterrence relying on kinetic military forces, whether they be conventional or nuclear.

domain-specific theory of cyber deterrence. Let me begin by focusing on the key elements of deterrence theory.

2 Deterrence Theory

Deterrence is a coercive strategy based on conditional threats with the goal of persuading the opponent to behave in a desirable way. It encompasses elements of control and power (both political and military) and usually has a medium- and long-term impact on the international arena. While one may trace the debate on deterrence strategies back to the 1920s and 1930s, deterrence rose to prominence only in the aftermath of World War II, when military power went from being a means to defeat the adversary, or at least of making the adversary's victory more costly than planned, to being considered as a key piece of bargaining power employed to avoid wars by means of coercion and intimidation (Possony 1946; Schelling and Affairs 1966; Schelling 1980; Brodie 1978; Zagare and Kilgour 2000; Powell 2008). It was this shift in the understanding of military power that made deterrence possible and a particularly valuable tool in avoiding nuclear conflicts.

It follows that most of the existing analyses on deterrence have focused on East–West nuclear tensions, in particular on policies defined between the late 1940s and the 1990s to deter nuclear attacks. These analyses assumed the bipolar scenario (USA vs Soviet Union) within which deterrence seemed the obvious approach to avoid conflicts and did not focus on

> how strategic relationship of this sort might come to be established in the first place when the core [problem] was that it existed and somehow it had to be survived (Freedman 2004, 22).

Freedman's words capture the *pragmatism* of deterrence theory, which rests on three elements: (i) a context in which actors, political dynamics, interests, and military and strategic options are clearly defined; (ii) the urgency of defining effective strategies that are immediately deployable in order to avoid a nuclear conflict; both leading to (iii) deterrence theory being tantamount to deterrence policies. Indeed, the so-called three waves of deterrence theory (Jervis 1988) actually identify different policy approaches endorsed by policy-makers and decision-makers between the late 1940s and the 1990s, rather than different theoretical stances on deterrence.

More in details, the first wave stems from Brodie's analysis of power and is based on the assumption that nuclear power was ever to be threatened and never to be deployed (Brodie 1978). The increasing reliance on rational choice theory to maximise the bargaining of power and ensure stability characterised the second wave (Powell 2008). The third wave arose in the 1980s (Jervis 1979) and led to the dismissing of deterrence theory in international relations as a theory that was hampering, rather than encouraging, a peaceful conclusion of the Cold War. The first two 'waves of deterrence' characterise deterrence theory and will be the focus of this section.

First and second waves deterrence strategies are modelled as follows: A believes that B is planning to attack it. In order to avoid the attack, A makes an explicit commitment to take action against B, should B decide to attack. A's commitment should be such that

B is convinced that any action against A will fail, because A has the capacity either to resist or punish B, and to outweigh any prospective gains for B. B's conviction hinges on A's signalling and credibility to act as it threatens. According to this model, we find here the three core elements of deterrence theory: the identification of the opponent (attribution); defence and retaliation as types of deterrence strategies; and the capability of the defender to signal credible threats (see Fig. 1).

This is a minimalist model of international deterrence (D_M) defined according to deterrence theory. The D_M model is defined at a high level of abstraction (LoA; Floridi 2008) and disregards the dynamics and characteristics of specific scenarios. It assumes rational agents (a minimal assumption, given states are expected to act rationally) but it does not depend on the kinds of weapons (nuclear or conventional), the kinds of relationships between the opponents (symmetric or non-symmetric), the levels of interaction between A and B (diplomatic or not), and the scope (general or tailored) of deterrence. One may enrich the model with information about these aspects. More details would make it more complex but would not change the dynamics and the elements identified in the D_M model.

As shown in Fig. 1, the three core elements of the model are intertwined. Attribution is essential for deterrence, as it allows the defender to identify the target of its strategy, and also conveys a credible signal to the right opponent. At the same time, conveying a credible, coercive message to (try to) change the offender's behaviour is key in any deterrence dynamic (Libicki 2009; Bunn 2007; Jensen 2012). Indeed, effective deterrence hinges on the defender signalling its intention to use its capabilities against the offender. Credible signalling exists in a relation of mutual dependence with the deterrence strategies. That is, the chosen type of deterrence determines and underpins the content of the message and its credibility, while signalling is crucial to exploit and convey the deterring capabilities of defender, whether they be to defend or retaliate.

Of the three elements identified in the D_M model, attribution and credible signalling are not controversial.[2] The identification of defence and retaliation as the two fundamental types of deterrence strategies may be more problematic, as it may be criticised for being too limited, and thus undermining the completeness of the D_M model. One may claim that the model should be expanded to include other deterrence strategies, which (at a first glance) do not rely on defence or retaliation, for example, deterrence by association, by norms and taboos, and by entanglement. This would be a mistake. Deterrence by association, by norms and taboos, and by entanglement differs from deterrence by retaliation only at first glance. A more attentive analysis reveals that they are simply different instances of deterrence by retaliation. One should think of them as *tokens* of the same *type* of deterrence strategy, insofar as

> each [strategy] occurs in a slightly different way, but all seek to punish and curb behaviour by adding a social cost (Ryan 2017).

[2] Attribution may not be necessary in all instances of deterrence, for example for deterrence by defence. Some argue that when the exact source of an attack is unknown, attribution and, hence, responsibility for an attack can be shifted to the particular state in which the attack originated (Morgan 2010; Goodman 2010). However, clear attribution remains necessary for deterrence by retaliation.

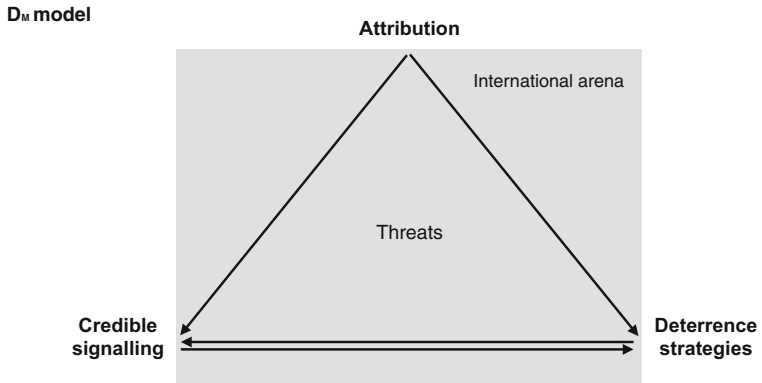

Fig. 1 The minimalist model of international deterrence (D_M) and the dependences among its elements

By adopting the D_M model, one does not deny that there are many different ways to implement deterrence; simply one regards these as tokens of the two fundamental types of strategies: defence and retaliation. Simply, the D_M model focuses on types rather than on tokens (the reader may recall the high granularity of the LoA).

In the same vein, the model specifies attribution and signalling as essential for deterrence but does not account for the different ways in which attribution can be ascertained; nor does it distinguish among the many possible modes of communication between the aggressor and the defender. Indeed, a model focusing on these aspects would be a model of specific implementations of deterrence theory, rather than a model of the theory itself.

By endorsing a high-granularity LoA, the D_M model can disregard the peculiarities (while not denying their existence) of specific cases and can focus on the necessary and sufficient elements of deterrence strategies as defined in deterrence theory. The extent to which the D_M model is valid in cyberspace will be indicative of the extent to which deterrence theory can be applied to this domain, while its limits indicate the problems that theory of cyber deterrence will have to address. Let us now consider in more detail the elements of the D_M model, starting with attribution.

3 Attribution

Attribution is crucial both for legal and strategic reasons. Legally, attribution helps the defender to legitimise its decision to retaliate (Clark and Landau 2011; Sterner 2011). Strategically, a correct and positive attribution underpins the coercive element of deterrence, as it directs retaliation against the actual offender ensuring that the right opponent is contrasted (Iasiello 2014). If attribution is dubious or mistaken, there are chances that the defender could attack the wrong opponent, prompting new frictions and conflicts.

At the same time, uncertain attribution weakens the logic of deterrence, as it impacts the cost–benefit analysis, which underpins deterrence strategies (Libicki 2009). In particular, from the perspective of the attacker, low chances to be identified make attacks appealing and strategically advantageous and undermine the threat of subsequent retaliation, as well as the credibility of the defender. In the eyes of the aggressor,

> [the] continuing inability to attribute attacks is tantamount to an open invitation [to attack] (Lan et al. 2010, 5).

Uncertainty of attribution also heightens the risk that retaliation may be perceived either as a mistaken response or as an escalation and, hence, may spark new frictions and conflicts, defeating the very purpose of deterrence. As Libicki stresses, in deterrence

> the lower the odds of getting caught, the higher the penalty required to convince potential attackers that what they might achieve is not worth the cost. Unfortunately, the higher the penalty […], the greater the odds that the [retaliation] will be viewed as disproportionate—at least by third parties and perhaps even by the attacker (Libicki 2009, 43).

These are problems faced by deterrence in cyberspace, where attribution is at best problematic, if not impossible (Libicki 2009; Goodman 2010; Jensen 2012; Haley 2013). For example, Jensen reports that most cyber attacks up until 2011 remain unattributed (Jensen 2012) and recent reports show that things are no different now.[3]

Dubious attribution is a consequence of both the distributed nature of both cyberspace, which facilitates anonymity, and of the way cyber attacks are conducted. These attacks are often launched in different stages and involve globally distributed networks of machines, as well as pieces of code that combine different elements provided (or stolen) by a number of actors. This was the case, for example, of NotPetya, the malware used for an allegedly state-run cyber attack,[4] which combines a vulnerability (EternalBlue) stored by the US National Security Agency (NSA) with an ordinary remote management tool (PsExec) to access computers, gain control, and extract relevant information, such as login credentials.[5] NotPetya has inflicted serious damage worldwide and, despite recent investigations linking the attack to North Korea,[6] attribution could not be proved positively, precisely because of the use of different tools and the particular dynamics of the attack.

In this scenario, identifying the malware, the network of infected machines, or even the country of origin of the attack is not sufficient for attribution, as it is well known that attackers can design and route their operations through third-party machines and countries with the goal of obscuring or misdirecting attribution. This leads some to maintain that uncertainty of attribution is inherent to the nature and the dynamics of cyberspace and that to solve it, we need to reengineer the Internet (Kastenberg 2009; Hollis 2011). This is the view, for example, of the former director of the NSA, John Michael McConnell:

> we need to reengineer the Internet to make attribution, geolocation, intelligence analysis and impact assessment—who did it, from where, why and what was the result—more manageable (McConnell 2010).

[3] http://www.nato.int/docu/review/2013/Cyber/timeline/EN/index.htm
[4] http://www.wired.co.uk/article/petya-malware-ransomware-attack-outbreak-june-2017
[5] https://www.theregister.co.uk/2017/06/28/petya_notpetya_ransomware/
[6] http://www.telegraph.co.uk/technology/2017/05/23/highly-likely-wannacry-cyber-attack-linked-north-korea/

Others have considered a different approach, stressing that attribution is not binary but comes in degrees of certainty (Jensen 2009, 2012). Building on this view, for example, Haley (Haley 2013) identifies ten degrees of certainty with which an attack can be attributed and proposes a 'spectrum of state responsibility' indicating different state responses (ranging from ignoring the attack to counterattack) depending on the degree of certainty of attribution. While this approach offers some guidance to policy-makers having to deal with dubious attribution of minor, non-kinetic cyber attacks (Iasiello 2014), it does leave open the question of state retaliation to severe cyber attacks whose source has not been identified positively, as well as national strategies to deter future severe attacks from unknown sources.

To be effective and incontestable, deterrence must be certain, severe, and immediate. Prompt, positive attribution is crucial to this end: the less positive the attribution, the more time will be necessary to respond; the less immediate is attribution, the less severe will be the defender's response. Hence, the limits of attribution in cyberspace pose serious obstacles to the deployment of effective deterrence strategies informed by deterrence theory. Recalling the D_M model, without attribution deterrence cannot function, for defence and retaliation (and indeed, signalling) are left without a target and are undermined by the inability of the defender to identify the attacker.

One could object that attribution problems are technical and therefore indicative of flaws in the design of cyberspace itself rather than shortcomings in deterrence theory per se, and that, as such, they require technical, not theoretical, solutions. According to this objection, once the appropriate technical solutions are in place, the problem of attributing cyber attacks will no longer be an obstacle to the application of deterrence theory in cyberspace.

This objection is misguided, as it fails to grasp that uncertainty of attribution is a function of the ambiguity in cyber conflicts. Ambiguity is a constitutive feature of these kinds of conflicts. It characterises attribution but also the level of violence, the assessment of their impact, and the nature of the involved actors and targets (Dipert 2013; Taddeo 2014). Thus, ambiguity should not be regarded as an unfortunate and undesired aspect of cyberspace, which nation states will want to 'correct'. Quite to the contrary, ambiguity is a desired feature of cyber conflicts, and technologies and practices will be developed to maintain it. This is clear when considering, for example, the reluctance of state actors to define international norms for the use of state-run cyber attacks or the failure of the UN Group of Governmental Experts, tasked with providing recommendations to the General Assembly on how to regulate state conduct in cyberspace. A working theory of deterrence in cyberspace has to be able to account for the ambiguity inherent to cyber conflicts, rather than simply trying to circumvent or ignore it.

At the same time, as we shall see in the rest of this article, when applied in cyberspace, all three core elements of deterrence theory identified in the D_M model face serious problems, indicating that it would be problematic to design deterrence strategies in cyberspace that rely on this theory, even if attribution was not a problem. Deterrence theory simply does not account properly for the dynamics of cyber conflicts, the nature of cyberspace, and the malleability of cyber weapons (Taddeo 2017a, 2017b). The next section will focus on each of these aspects.

4 Defence and Retaliation

Both deterrence by defence and by retaliation include elements of coercion and control, although to different extents. Deterrence by defence is essentially concerned with controlling the impact of an attack either by preventing (dismounting) it or by rendering it ineffective (ensuring that, even if it breaches the defences, the attack does not reach its intended target). Both aspects act as deterrent as they ensure that new attacks will inevitably fail. Effective defence has also a coercive element; insofar as by discouraging or thwarting an otherwise successful attack, the defendant forces the opponent to change its behaviour.

Deterrence by retaliation is essentially coercive. It rests on the (threat of) use of force to change the offensive plan of the opponent. State A launches, or threatens to launch, a counter strike that imposes a cost on State B and outweighs B's benefit from the initial attack. In view of these likely costs, State B decides not to attack. Retaliation also has an element of control, as full control of the impact and the scope of the retaliation is crucial to avoid breaches of proportionality and risks of escalation.

Deterrence strategies based on either defence or retaliation (or a combination of both) are problematic, if not ineffective, when deployed in cyberspace.

4.1 Defence in Cyberspace

Anticipating, defence is guaranteed to be ineffective as a deterrence strategy in cyberspace because cyber defence mechanisms have little control over cyber attacks. This deprives defence of any strategic power (Taddeo 2017a, 2017b) and transforms it into a means of ensuring resilience of information systems, rather than a means to deter new attacks. Let me unpack this analysis.

Defence in cyberspace is porous in nature (Morgan 2010); every system has its security vulnerabilities and identifying and exploiting them is simply a matter of time, means, and determination. This makes ephemeral even the most sophisticated defence mechanisms, thus limiting their potential to deter new attacks (Taddeo 2017a). At the same time, even when successful, cyber defence does not lead to a strategic advantage, insofar as dismounting a cyber attack very rarely leads to the ultimate defeating of an adversary. This creates an environment of *persistent offence* (Harknett and Goldman 2016), where attacking is tactically and strategically more advantageous than defending.

An 'offence-persistent' environment differs from an 'offence-dominant' environment, in that defence is under constant stress, but is not superfluous, and the success of the offence is not a given. In an offence-persistent environment, defence can achieve tactical and operational success in the short term if it can adjust constantly to the means of attack, but it cannot win strategically. Offence will persist and the interactions with the enemy will remain constant (Harknett and Goldman 2016). In this kind of environment, deterrence by defence is guaranteed to be ineffective, as defence does not discourage attackers from their intention to offend. This is even more so in cyberspace, where uncertainty of attribution, low entry cost of attacks, and the inherently vulnerable nature of information systems encourage attackers to test defences.

In cyberspace, defence remains salient and necessary, but primarily as a means to guarantee the resilience of a system once an attack has been launched (and also after it

has breached the system), rather than as a means of deterring attackers (Bologna et al., 2013; Bendiek and Metzger 2015). Cyber defence, then, is more akin to safety engineering, in that it mitigates and manages the risk following an attack (Libicki 1997; Rattray 2009), rather than avoiding them.

4.2 Retaliation in Cyberspace

Given the guaranteed ineffectiveness of defence as a deterrence strategy, and offence persistency in cyberspace, state actors focus their attention on developing cyber deterrence by retaliation. As Croston stresses,

> the goal for major powers should not be the futile hope of developing a perfect defensive system of cyber deterrence, but rather the ability to instil deterrence based on a mutually shared fear of an offensive threat […] (Crosston 2011).

Approaches to deterrence by retaliation in cyberspace often make reference to nuclear deterrence models. Some consider mutual assured destruction (MAD) a viable strategy to shape cyber deterrence, given its potential to limit the freedom of major political actors to attack each other:

> By capitalizing on this shared vulnerability to attack and propagandizing the open build-up of offensive capabilities, there would arguably be a greater system of cyber deterrence keeping the virtual commons safe (Crosston 2011).

This approach rests on the idea that analyses and practices of nuclear deterrence can shed light on cyber deterrence (Owens et al. 2009). Nye outlines this idea quite clearly:

> There are some important nuclear-cyber strategic rhymes, such as the superiority of offense over defense, the potential use of weapons for both tactical and strategic purposes, the possibility of first- and second-use scenarios, the possibility of creating automated responses when time is short, the likelihood of unintended consequences and cascading effects […] (Nye 2011, 22–23).

Aside from the guaranteed ineffectiveness of defence as a deterrent, which is indeed an aspect peculiar to both nuclear and cyber conflicts, the rest of the similarities listed by Nye are too generic to characterise nuclear and cyber conflicts as equivalent. They can, actually, be used to describe many types of modern warfare; air and marine warfare, for example, meet all the requirements he lists.

An attentive analysis unveils that nuclear and cyber conflicts differ radically in several crucial aspects. Differences range from clarity of attribution, the destructive power of the attacks, the military capabilities of the opponents, and the nature of the involved actors (Sterner 2011; Taddeo 2016). As Libicki stresses

> [i]n the Cold War nuclear realm, attribution of attack was not a problem; the prospect of battle damage was clear; the 1,000th bomb could be as powerful as the first; counterforce was possible; there were no third parties to worry about;

> private firms were not expected to defend themselves; any hostile nuclear use crossed an acknowledged threshold; no higher levels of war existed; and both sides always had a lot to lose (Morgan 2003; Libicki 2009; Stevens 2012).

These differences shape diverging deterrence strategies. Nuclear deterrence is singular and symmetric (Libicki 2009), while cyber deterrence is repeatable and non-symmetric. Nuclear deterrence is singular; as by the time a nuclear attack and retaliation have run their course, both parties are likely destroyed and there is no chance for the offender to counter retaliate. At the same time, nuclear deterrence works only among actors enjoying symmetric military power: a state with no nuclear capacity could not deter a nuclear power on those terms.

Unlike nuclear deterrence, cyber deterrence is repeatable, as non-kinetic retaliations are unlikely to defeat the opponent definitively, let alone pose ultimate threats (Libicki 2009; 2017a), thus leaving the aggressor able to counter retaliate and favouring multiple interactions between defender and offender. Early analyses (Libicki 2009) maintain that cyber deterrence between states is symmetric, as it occurs among peers and the defender and offender are assumed to share the same strategic ground. This is only partially correct, as this view overlooks more complex scenarios, where the defender may have inferior cyber capabilities and may use (proportionate) kinetic means to retaliate, or where the offender relies on cyber means to attack an opponent with superior kinetic means. This is the case that Geers describes:

> [b]ecause cyber warfare is unconventional and asymmetric warfare, nations weak in conventional military power are also likely to invest in it as a way to off-set conventional disadvantages (Geers 2012, 5).

The non-symmetric use of cyber capabilities has also been acknowledged in a leaked NSA report (NSA 2013), which recognises that

> [c]yberattacks offer a means for potential adversaries to overcome overwhelming U.S. advantages in conventional military power and to do so in ways that are instantaneously and exceedingly hard to trace […] (NSA 2013, 3).

Even when focusing only on state actors, it is not possible to assume symmetry between the cyber capabilities of the defender and offender. For this reason, I argue that cyber deterrence is non-symmetric, as the defender and the offender may or may not enjoy the same kinetic and cyber capabilities. This is crucial, as it means than in deciding whether or not to retaliate, the defender will have to consider the possibility of both kinetic and non-kinetic counter retaliation and, hence, of escalation. The two couples 'singular and symmetric' and 'repeatable and non-symmetric' indicate that nuclear and cyber deterrence are not related and, thus, that analogies between nuclear and cyber deterrence are not warranted.

Deterrence strategies are also heavily determined by the nature of the threats that they pose. Nuclear conflicts pose existential threats—a nuclear attack is likely to destroy both opponents, their infrastructures, and populations; while cyber conflicts do not. Indeed, cyber conflicts proliferate because they pose *non-existential* threats

(Taddeo 2012; Floridi and Taddeo 2014). This difference is critical when considering deterrence. In nuclear deterrence, the existential nature of the threats justifies and makes credible MAD strategies. In contrast, cyber deterrence affords to the defender a whole range of possible strategies—from in-kind retaliation and economic sanctions to diplomatic measures and proportionate kinetic responses—because of the non-existential nature of (non-kinetic) cyber threats. These options are lost when modelling cyber deterrence in analogy to nuclear deterrence.

4.3 Control and Risks of Cyber Deterrence by Retaliation

Even when not informed by analogies with nuclear strategies, retaliation as identified in deterrence theory raises serious problems when applied in cyberspace. Unlike with defence, deterrence by retaliation is not guaranteed to be ineffective in this domain. Indeed, in an offence-persistent environment like cyberspace, retaliation can actually be a successful strategy. However, when deployed in cyberspace, the nature of cyber weapons and of cyber conflicts undermines the control element of retaliation, making it a hazardous choice for deterrence.

Retaliation is coupled with the risk of escalation. This risk is amplified when retaliation occurs in a non-symmetric scenario, where the opponent may lack cyber capabilities and instead counter-retaliates using kinetic means. Control over the weapons used and the impact of the resulting retaliation is crucial to avoid escalation. In cyberspace, however, this control is limited given the *malleability* of cyber weapons.

Cyber weapons are malleable insofar as they can be accessed, stored, combined, repurposed, and redeployed much more easily than was ever possible with other kinds of military capability (Schneier 2017). Repurposing or redeploying state-designed or state-owned malware is not too rare an event. It happened in 2011 with Stuxnet. Despite being designed to target specific configuration requirements of Siemens software installed on Iranian nuclear centrifuges, the worm was eventually released on the Internet and infected systems in Azerbaijan, Indonesia, India, Pakistan, and the USA.[7] Even more worryingly, the vulnerability that Stuxnet exploited has been used for at least 6 years to weaponise Angler, one of the most infectious malwares used by cyber criminals to target online banking websites.[8] In the same vein, in 2017, two major cyber attacks, WannaCry and NotPetya, repurposed an exploit (EternalBlue) stolen from the NSA.[9]

The chances of a cyber weapon causing more damage than originally planned increase when considering the ever more likely deployment of 'counter autonomy' systems for national defence. These are machine-learning systems that are able to identify and target autonomously vulnerabilities in other systems, while also isolating and patching their own.[10] As machine-learning systems learn and evolve in an unsupervised way, their use for defence purposes poses concrete risks of unforeseen, disproportionate damage (Cath et al. 2017).

[7] https://www.symantec.com/security_response/writeup.jsp?docid=2010-071400-3123-99
[8] https://www.theregister.co.uk/2016/05/09/sixyearold_patched_stuxnet_hole_still_the_webs_biggest_killer/
[9] https://www.forbes.com/sites/thomasbrewster/2017/05/12/nsa-exploit-used-by-wannacry-ransomware-in-global-explosion/#3f04a279e599
[10] https://fas.org/irp/agency/dod/dsb/autonomy-ss.pdf; https://www.darpa.mil/program/cyber-grand-challenge

The malleability of cyber weapons erodes the control element of retaliation in cyberspace and, in so doing, makes retaliation a dangerous strategic choice, with the potential for disastrous cascade effects. Weak control over the impact of retaliation could lead to a breaching of proportionality, in turn triggering self-defence and prompting escalation. Ensuring control over retaliation is essential to avoid these unintended effects and respecting proportionality is crucial to this end. As Iasiello put it:

> A nation state must not only strike back against the aggressor but it must do so in a way as to make its point—that is, it must be a forceful strike—but not so forceful as to solicit negative reaction in the global community […] (Iasiello 2014, 59).

There is a general consensus that the principle of proportionality applies in the case of cyber deterrence and that it does not require an in-kind response (Libicki 2009; Jensen 2009; Goodman 2010; Iasiello 2014). Hence, a consensus that retaliation to a cyber attack could use cyber or kinetic means (or a combination of them), as long as the response is comparable to the impact of the initial attack and does not equate to an escalation (Hathaway and Crootof 2012).

However, determining the impact of a non-kinetic cyber attack is problematic. Proportionality prescribes that retaliation should equate to the actual (and not just the discovered) damage suffered by the defender. This can be a serious hurdle in cyberspace,

> [where] very little […] can be inferred about unseen activities (which cannot be measured) from those that are seen (which can be measured) (Libicki 2009, 103).

At the same time, even when the attack is detected and its impact is clear, it can be difficult to assess the value and type of damage, and therefore an appropriate response. As Harknett and Goldman note:

> If an attack reduces no buildings to rubble and kills no one directly, but destroys information, what is the response? We tend to think about information as intangible, but the loss of information can have tangible personal, institutional, and societal costs. What credibly can be placed at risk that would dissuade a state from contemplating such an attack? (Harknett and Goldman 2016).

The questions that they pose hinge on what I have described in a previous article (Taddeo 2014) as an *ontological gap* between Just War Theory and cyber conflicts. This gap refers to the difference between the ontology assumed by Just War Theory—which is centred on human beings, tangible objects, and kinetic conflicts causing physical damage and bloodshed—and the nature of (artificial) agents, (digital) targets, and (non-kinetic) cyber conflicts (Taddeo 2012). Because of this gap, it is problematic to apply Just War Theory principles to cyber conflicts. In the case of proportionality and deterrence, in particular, we face the danger that uncertainties in assessing proportionality may justify self-defence and facilitate escalation.

While proportionality still remains a valid and desirable principle to regulate cyber conflicts, its application poses serious problems that can only be resolved once the ontological gap is overcome. In turn, addressing the gap will require an ethical framework for the regulation of cyber conflicts that takes account of the moral stance of informational entities, like artificial agents and digital targets, as well as that of human beings and tangible objects (Taddeo 2012; Floridi 2013; Floridi and Taddeo 2014). Without such a framework, attempts to define deterrence strategies in cyberspace that are able to respect and make sense of proportionality are doomed to failure.

In kinetic scenarios, defence and retaliation strategies offer the perfect balance between control of response and coercion that ultimately allows the defender to show its power and deter the offender. In cyberspace, this balance is not achievable, as both defence and retaliation lack control. Recalling the D_M model, neither of these two types of deterrence strategies works in cyberspace. However, there is an important difference to be noted: while defence strategies are guaranteed to be ineffective in an offence-persistent environment, like cyberspace, retaliation could be a viable strategic choice (within the limits posed by attribution).

Nonetheless, to be successful, retaliation needs to be reconsidered to ensure that, while remaining essentially about coercion, it can rely on strong control mechanisms that ensure a proportionate response. This requires addressing the ontological gap and overcoming the limits of Just War Theory in cyberspace. The alternative is to model retaliation in cyberspace using MAD strategies, but this is more likely to lead to escalation than it is to deter new conflicts.

5 Credible Signalling

A defender deters prospective attackers by signalling to the attacker its awareness of the offender's plans and the envisaged response, should the plan be implemented. Without this signalling, deterrence would not be possible. Iasiello, for example, notes that retaliation becomes ineffective and can be misinterpreted if the defender is not able to convey a credible signal of its intentions (Iasiello 2014).

As shown in the D_M model, signalling is only effective insofar as it conveys a coercive message (threat) and, thus, it depends on the deployment of an appropriate deterrence strategy (see Fig. 1). The message has to be credible. The credibility of the message hinges on the reputation of the defender to follow through on its threats (Freedman 2004). Indeed, reputation is a central aspect of deterrence theory. Famously, Schelling stressed that

> 'Face' is one of few things worth fighting over […] 'face' is merely the interdependence of a country's commitments; it is a country's reputation for actions, the expectations other countries have about its beahviour (Freedman 2004, 53).

In kinetic scenarios, reputation is gained by showcasing a state's military capabilities—military parades and deployment of soldiers or ships on the borders of the offending state typically serve this purpose—as well as by showing ability to resolve (to deter or defeat the opponent) over time. To some extent, the same also

holds true in cyberspace, where a state's reputation also refers to a state's past interactions in this domain, its known cyber capabilities to defend and offend, as well as its overall reputation in resolving conflicts. One caveat is that a state's reputation in cyberspace may not necessarily correspond to its actual capabilities in this domain, as states are reluctant to circulate information about the attacks that they receive. In the medium and long terms, this may make signalling less credible and thus more problematic, than in other domains of warfare.

Signalling can be either general or tailored. General signalling conveys a message about the overall deterrence strategy to the rest of the international arena, through open statements released by a state conveying information about its approaches, commitments, and capabilities. Although it may be problematic in some circumstances, general signalling in cyberspace is not impossible. The reference to the ability to resort to the 'full range of tools available to the United States' in the US cyber strategy document (US Government 2015, 14) as well as mention of the Active Cyber Defence capabilities in the UK equivalent (UK Government 2015) serves precisely this purpose. In both cases, general signalling is credible as it rests on the reputation that the USA and UK have in cyberspace, as well as in the international arena.

Tailored signalling—the conveying of a threat to a specific offender indicating the possible targets of retaliation—is more problematic than general signalling and constitutes a significant obstacle to delivering effective deterrence strategies in cyberspace. This kind of signalling is effective if attribution is certain. If the defender has not identified the offender correctly, tailored signalling can be counterproductive given it may be directed to the wrong actor. Tailored signalling also requires a careful fine-tuning in order not to expose the defender's capabilities and assets, especially when the defender is considering retaliation in-kind.

The risks are multiple and range from exposing knowledge about the opponent's cyber assets, which would imply that the defender has also run cyber operations (sabotage or espionage) against the opponent, to revealing the defender's assets and strategies, which may expose and therefore render futile its cyber capabilities, such as zero-day exploits (for example). At the same time, too vague a signalling would undermine the credibility of the threats and the success of deterrence. Two alternatives follow for cyber deterrence: signalling could become increasingly decoupled from reputation, though thereby also weakening the coercive nature of deterrence; or deterrence could become less about signalling as a way of alerting the opponent and more about *demonstrating* the capabilities and intentions of the defender. But this will pose increasing risks for escalation. Both scenarios undermine the chances of successful deterrence strategies.

6 Conclusion

The limits of deterrence theory in cyberspace are not trivial and they follow the fundamental differences between kinetic and cyber conflicts outlined in this article. These differences cannot be disregarded when defining deterrence strategies in cyberspace. As have I argued elsewhere (Taddeo 2017b), understanding these differences—and identifying their impact on international relations and on military strategies—is a preliminary and necessary step to any attempt to develop strategies for cyber deterrence. For this reason, we must recognise

the limits of approaching cyber deterrence by analogy with kinetic conflicts and move past them. As Betz and Stevens put it:

> It is little wonder that we attempt to classify […] the unfamiliar present and unknowable future in terms of a more familiar past, but we should remain mindful of the limitations of analogical reasoning in cyber security (Betz and Stevens 2013).

Analogies can be powerful for they inform the way in which we think and constrain ideas and reasoning within a conceptual space (Wittgenstein 2009). However, if the conceptual space is not the right one, analogies become misleading and detrimental for any attempt to develop innovative and in-depth understanding of new phenomena, and they should be abandoned altogether. When the conceptual space is the right one, analogies are at best a step on Wittgenstein's ladder and need to be left behind once they have taken us to the next level of the analysis. In the best scenario, this is the case of the analogies between traditional and cyber deterrence.

Once we have abandoned any unhelpful analogies with traditional deterrence, a theory of cyber deterrence will have to develop an original approach to address the specificities of cyber conflicts. These span three areas: conceptual, normative, and regulative.

Conceptually, the non-kinetic nature of cyber conflicts has redefined our understanding of key notions such as harm, violence, target, combatants, weapons, attack, and political power, and a theory of cyber deterrence will be successful only insofar as it rests on a clear grasp of these concepts. At the same time, the ontological gap and the limits of Just War Theory, identified in section 4.3, highlight the absence of clear ethical guidance in shaping cyber deterrence. There is a pressing need to define ethical principles that would ensure the deployment of deterrence strategies able to respect individual rights (Taddeo 2013; Taddeo and Glorioso 2016), avoid unnecessary violence and bloodshed, and ultimately foster a peaceful geopolitical environment. Deterrence must be deployed in accordance with international humanitarian laws, and conceptual and normative problems pose sever hindrances to the regulation of cyber deterrence. However, as these laws were defined with respect to kinetic conflicts, it is unclear to what extent, if at all, they offer the right guidance to regulate cyber deterrence (reference removed for double-blid review), and therefore to guarantee the stability of international relations of current and future information societies.

Understanding how deterrence in cyberspace can be deployed while addressing the conceptual, normative, and regulative challenges will be the goal of my future work.

References

Bendiek, Annegret, and Tobias Metzger. (2015). Deterrence theory in the cyber-century: lessons from a state-of-the-art literature review lecture notes in informatics (LNI), Gesellschaft Fur Informatik, Bonn, 2015. *Lecture Notes in Informatics*, 553–70.

Betz, D. J., & Stevens, T. (2013). Analogical reasoning and cyber security. *Security Dialogue, 44*(2), 147–164. https://doi.org/10.1177/0967010613478323.

Bologna, S., Fasani, A., & Martellini, M. (2013). From fortress to resilience. In M. Martellini (Ed.), *Cyber security: deterrence and IT protection for critical infrastructures* (pp. 53–56). Heidelberg: Springer.

Brodie, B. (1978). The development of nuclear strategy. *International Security, 2*(4), 65–68.

Bunn, M. E. (2007). *Can deterrence be tailored? Strategic Forum, Number 225, January 2007*. Washington, D.C, USA: Institute for National Strategic Studies, National Defense University.

Cath, C., Wachter, S., Taddeo, M., & Floridi, L. (2017). Artificial intelligence and the "Good Society": the US, EU, and UK approach. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-017-9901-7.

Clark, D., & Landau, S. (2011). Untangling attribution. *Harvard National Security Journal, 2011*(2), 25–40.

Crosston, M. (2011). World gone cyber MAD: how "Mutually Assured Debilitation" is the best hope for cyber deterrence. *Strategic Studies Quarterly, 50*(1), 100–116.

Dipert, Randall. 2013. The essential features of an ontology for cyberwarfare. In *Conflict and Cooperation in Cyberspace*, edited by Panayotis Yannakogeorgos and Adam Lowther, 35–48. Taylor & Francis. http://www.crcnetbase.com/doi/abs/10.1201/b15253-7.

European Union. (2014). 'Cyber defence in the EU: preparing for cyber warfare?—think tank'. Brussels. http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2014)542143.

European Union. (2015). Cyber diplomacy: EU dialogue with third countries—think tank'. Brussels. http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2015)564374.

Floridi, L. (2008). The method of levels of abstraction. *Minds and Machines, 18*(3), 303–329. https://doi.org/10.1007/s11023-008-9113-7.

Floridi, L. (2013). *The ethics of information*. Oxford: Oxford University Press.

Floridi, L., & Taddeo, M. (Eds.). (2014). *The ethics of information warfare. Law, Governance and Technology Series* (Vol. 14). Heidelberg: Springer.

Freedman, L. (2004). *Deterrence*. Cambridge: Polity Press.

Geers, K. (2012). *Sun Tzu and cyber war*. Tallinn: Cooperative Cyber Defence Centre of Excellence.

Goodman, Will. (2010). 'Will Goodman, Cyber deterrence: tougher in theory than in practice?, STRATEGIC STUD. Q., Fall 2010, at 102, 102'. *TRATEGIC STUD. Q.*, 2010 (Fall): 102–35.

Haggard S., & Simmons, B. A. (1987) Theories of international regimes. International Organization 41 (03): 491

Haley, Cristopher. 2013. 'A theory of cyber deterrence'. *Georgetown Journal of International Affairs* February. http://journal.georgetown.edu/a-theory-of-cyber-deterrence-christopher-haley/.

Harknett, R. J., & Goldman, E. O. (2016). The search for cyber fundamental. *Journal of Information Warfare, 15*(2), 81–88.

Hathaway, O., & Crootof, R. (2012). The law of cyber-attack. *California Law Review, 100*(1–2012), 817–886.

Hollis, D.B. 2011. An E-SOS for cyberspace. HARV. INT'L L.J. *52*(373), 374–75.

Iasiello, E. (2014). Is cyber deterrence an illusory course of action? *Journal of Strategic Security, 7*(1), 54–67.

International Security Advisory Board. (2014). A framework for international cyber stability. United States Department of State. http://goo.gl/azdM0B.

Jensen, E. T. (2009). Cyber warfare and precautions against the effects of attacks. *Texas Law Review, 88*(1533), 1534–1569.

Jensen, Eric Talbot. 2012. Cyber deterrence. SSRN Scholarly Paper ID 2070438. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2070438.

Jervis, R. (1979). Deterrence theory revisited. *World Politics, 31*(2), 289–324. https://doi.org/10.2307/2009945.

Jervis, R. (1988). Realism, game theory, and cooperation. *World Politics, 40*(3), 317–349. https://doi.org/10.2307/2010216.

Kastenberg, J. E. (2009). Changing the paradigm of Internet access from government information systems: a solution to the need for the DoD to take time-sensitive action on the Niprnet. *Air Force Law Review, 64*, 175.

Kugler, R. (2009). Deterrence of cyber attacks. In F. Kramer, S. Starr, & L. Wentz (Eds.), *Cyberpower and national security* (pp. 309–342). Washington, D.C.: National Defense University.

Lan, Tang, Zhang Xin, Harry Raduege Jr., Dmitry Grigoriev, Pavan Duggal, and Stein Schjølberg. 2010. Global cyber deterrence views from China, the U.S., Russia, India, and Norway. EastWest Institute.

Libicki, Martin C. (1997). Defending cyberspace and other metaphors.

Libicki, Martin C. (2009). Cyber deterrence and cyberwar. Product Page. http://www.rand.org/pubs/monographs/MG877.html.

Libicki, Martin. 2011. "The Strategic Uses of Ambiguity in Cyberspace." Military and Strategic Affairs 3 (3): 3–10.

McConnell, Mike. 2010. Mike McConnell on how to win the cyber-war we're losing, February 28. http://www.washingtonpost.com/wp-dyn/content/article/2010/02/25/AR2010022502493.html.

Morgan, P. M. (2003). *Deterrence now. Cambridge Studies in International Relations 89. Cambridge [England]*. New York: Cambridge University Press.

Morgan, Patrick M. 2010. Applicability of traditional deterrence concepts and theory to the cyber realm. In *Proceedings of a Workshop on Deterring Cyberattacks: Informing Strategies and Developing Options for U.S. Policy*, 55–76. Washington, D.C, USA: National Academic Press.

NSA. (2013). A strategy for surveillance powers. *The New York Times*. http://www.nytimes.com/interactive/2013/11/23/us/politics/23nsa-sigint-strategy-document.html.

Nye, J. S. (2011). Nuclear lessons for cyber security? *Strategic Studies Quarterly, 5*(4), 11–38.

Owens, W. A., Dam, K. W., Lin, H., & National Research Council (U.S.), National Research Council (U.S.), and National Research Council (U.S.), eds. (2009). *Technology, policy, law, and ethics regarding U.S. acquisition and use of cyberattack capabilities*. Washington, DC: National Academies Press.

Possony, S. T. (1946). Atomic power and world order. *The Review of Politics, 8*(4), 533–535.

Powell, R. (2008). *Nuclear deterrence theory: the search for credibility*, *Digitally printed version. Paperback Re-Issue*. Cambridge: Cambridge University Press.

Rattray, G. J. (2009). An environmental approach to understanding cyberpower, in Kramer, Cited, 253-274, Esp. 256. In S. S. Kramer & L. K. Wentz (Eds.), *Cyberpower and national security* (pp. 253–274). Washington, D.C.: National Defense UP.

Ryan, N. J. (2017). Five kinds of cyber deterrence. *Philosophy & Technology*. https://doi.org/10.1007/s13347-016-0251-1.

Schelling, T. C. (1980). *The strategy of conflict: [with a new preface]*. Cambridge, Mass: Harvard Univ. Press.

Schelling, Thomas C., and Harvard University Center for International Affairs. 1966. Arms and influence. Yale University Press.

Schneier, Bruce. (2017). Why the NSA makes us more vulnerable to cyberattacks'. *Foreign Affairs*, May 30. https://www.foreignaffairs.com/articles/2017-05-30/why-nsa-makes-us-more-vulnerable-cyberattacks.

Sterner, E. (2011). Retaliatory deterrence in cyberspace. *Strategic Studies Quaterly, 5*(1), 65–80.

Stevens, T. (2012). A cyberwar of ideas? Deterrence and norms in cyberspace. *Contemporary Security Policy, 33*(1), 148–170. https://doi.org/10.1080/13523260.2012.659597.

Taddeo, M. (2012) Information Warfare: A Philosophical Perspective. Philosophy & Technology 25 (1):105-120

Taddeo, M. (2013) Cyber Security and Individual Rights, Striking the Right Balance. Philosophy & Technology 26 (4):353-356

Taddeo, M. (2014). 'Information Warfare: The Ontological and Regulatory Gap'. Newsletter on Philosophy and Computers 14 (1 (fall 2014)): 13–20.

Taddeo, M. (2016) On the Risks of Relying on Analogies to Understand Cyber Conflicts. Minds and Machines 26 (4):317-321

Taddeo, M. (2017a) Cyber Conflicts and Political Power in Information Societies. Minds and Machines 27 (2): 265-268

Taddeo, M. (2017b) Deterrence by Norms to Stop Interstate Cyber Attacks. Minds and Machines 27 (3):387-392

Taddeo, M., & Glorioso, L. (Eds.) (2016). 'Regulating Cyber Conflicts and Shaping Information Societies'. In Ethics and Policies for Cyber Operations. Philosophical Studies Series. Berlin, Heidelberg: SPRINGER.

Tanji, Michael. 2017. 'Deterring a cyber attack? Dream on...' *WIRED*. Accessed July 15. https://www.wired.com/2009/02/deterring-a-cyb/.

UK Government. (2014). Deterrence in the twenty-first century: government response to the Committee's Eleventh Report. http://www.publications.parliament.uk/pa/cm201415/cmselect/cmdfence/525/52504.htm.

UK Government. (2015). *National security strategy 2016–2021*. London: HM Government https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/567242/national_cyber_security_strategy_2016.pdf.

UN Institute for Disarmament Research. 2014. Cyber stability seminar 2014: preventing cyber conflict.

US Government. (2015). The Department of Defense cyber strategy. Washington, D.C, USA.

Wittgenstein, L. (2009). *Philosophical investigations* (Rev. 4th ed.). Chichester, West Sussex, U.K. ; Malden, MA: Wiley-Blackwell.

Zagare, F. C., & Marc Kilgour, D. (2000). *Perfect deterrence. Cambridge Studies in International Relations 72*. Cambridge: Cambridge University Press.