Philosophy & Technology (2018) 31:323–329
https://doi.org/10.1007/s13347-018-0328-0

BRIEF COMMUNICATION

Deterrence and Norms to Foster Stability in Cyberspace

Mariarosaria Taddeo [1,2,3]

Published online: 10 August 2018

Abstract

Deterrence in cyberspace is possible. But it requires an effort to develop a new domain-specific, conceptual, normative, and strategic framework. To be successful, cyber deterrence needs to shift from threatening to prevailing. I argue that by itself, deterrence is insufficient to ensure stability of cyberspace. An international regime of norms regulating state behaviour in cyberspace is necessary to complement cyber deterrence strategies and foster stability. Enforcing this regime requires an authority able to ensure States compliance with the norms at an international level, run investigations into suspected State-run (or Statesponsored) cyber operations to define attribution, expose breaches of the norms, and impose adequate sanctions and punishments. These requirements define a political mandate for an authority that will have a deep impact on international relations and geo-political equilibriums. The UN Security Council has the necessary resources and the political and coercive power to meet these requirements. The time has come to embrace this power to consolidate and enforce an international regime of norms to regulate state behaviour in cyberspace. Problems, mistakes, and even failures are to be expected, but they must not hinder the process.

Keywords Artificial intelligence · Cyber attacks · Deterrence theory · International relation · Stability

In March 2018, the US Computer Emergency Readiness Team (CERT) issued an alert on a series of cyber attacks attributed to Russian government and targeting US governmental offices and infrastructures in the energy, nuclear, water, aviation, and critical manufacturing sectors.[1] Later in the year, the director of the US National

[1] https://www.us-cert.gov/ncas/alerts/TA18-074A

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

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

[2] Department of Computer Science, University of Oxford, Oxford OX1 3QD, UK

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

Intelligence echoed the CERT's alert, stating that the US digital infrastructure "is literally under attack".[2] The US case offers a good example of the level of threat that cyber attacks pose to national security and defense of mature information societies (Floridi 2016).

Cyber attacks are escalating in frequency, impact, and sophistication. And state actors often play a central role in the escalation process. Starting in 2003, States have relayed frequently on cyber operations for espionage and sabotage purposes. Well-known examples range from Titan Rain (2003), the Russian attack against Estonia (2006) and Georgia (2008), to Stuxnet and Operation Olympic Game (2006–2012), WannaCray and NotPetya (2017). This trend will continue. The relatively low entry-cost and the high chances of success mean that states will keep developing, relying on, and deploying cyber attacks, thus increasing the risks of their escalation.

Scholars, militaries, and policy makers have stressed that deterrence may play a significant role in mitigating these risks and fostering stability of cyberspace (Freedberg 2014; UN Institute for Disarmament Research 2014; Taddeo 2017b). Most of the existing analyses refer to deterrence theory.[3] They address deterrence as a coercive strategy based on conditional threats with the goal of persuading the opponent to behave in a desirable way. According to this theory, in a scenario in which as State is planning to attack another State, deterrence will be effective if the defendant is able to identify with certainty the opponent (attribution) and communicate to it (signalling) a credible threat (punishments or denial) proportionate to the damage that the opponent is planning to cause, but severe enough to outweigh any advantage that the opponent may gain from attacking.

However, applying deterrence theory to cyberspace poses serious problems. The distributed and the interconnected nature of the domain (Chadwick and Howard 2009) makes it difficult to define territoriality and sovereignty of States and hence to identify the boundaries for States' actions. The non-physical nature of cyber attacks (Taddeo 2012) hampers the assessment of the damage that they may cause and, hence, of the proportionality of responses. The difficulties to attribute with certainty cyber attacks to their authors undermine the very core of deterrence theory: if the opponent cannot be identified, it is impossible to issue a meaningful threat.

The success of deterrence theory, and some suggest of deterrence itself, in cyberspace hinges upon the possibility to address these problems (Kugler 2009). Some analyses maintain that, given the differences with kinetic (violent) conflicts, solving these problems is impossible and that deterrence theory cannot be applied to the case of cyber conflicts (Lan et al. 2010). They conclude that deterrence in cyberspace is unattainable. Others hold the opposite view, and stress that it is possible to deter in cyberspace, precisely because deterrence theory can be successfully applied in this domain (Crosston 2011).

Both positions are misled. They both draw on an analogy between deterrence of cyber attacks and deterrence of kinetic attacks (Taddeo 2016) and conclude that cyber

[2] https://www.huffingtonpost.co.uk/entry/dan-coats-warns-of-dangerous-new-cyber-attacks_us_5b4aa5b5e4b0bc69a787c923?guccounter=1&guce_referrer_us=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_cs=u6PVg4rySjDpF-pC2f7q1g

[3] I shall refer to deterrence theory to indicate any theory of deterrence (in particular, first and second-wave theories) relying on kinetic military forces, whether conventional or nuclear (Brodie 1978; Powell 2008).

deterrence is possible only insofar as deterrence theory can be applied to cyberspace (Owens et al. 2009; Nye 2011). This approach overlooks the specific nature of cyber attacks and cyberspace, and disregards the dynamics of cyber conflicts (Taddeo (2014, 2016, 2017b).

Kinetic and cyber conflicts differ radically in several, crucial aspects, ranging from clarity of attribution, the destructive power of the attacks, to the nature of the involved actors and targets (Libicki 2009; Floridi and Taddeo 2014). For these reasons, analogies between cyber and kinetic conflicts are not warranted and should be abandoned. Efforts should focus on developing an in-depth understanding of cyberspace and cyber conflicts and define a domain-specific framework for deterrence. The alternative is risky. It is equivalent to forcing the proverbial square peg (cyber deterrence) into a round hole (deterrence theory): we are more likely to smash the toy than to win the game. As USN Commander Bebber stated:

> [Military] history suggests that applying the wrong operational framework to an emerging strategic environment is a recipe for failure. During the World War I, both sides failed to realize that large scale artillery barrages followed by massed infantry assaults were hopeless on a battlefield that strongly favored well-entrenched defense supported by machine gun technology. […] The failure to adapt had disastrous consequences.[4]

This is also the case when considering cyber deterrence. Analyses of cyber deterrence need to consider the strategic nature of cyberspace and the new capabilities availed by technology, in order to provide the right conceptual framework and define successful deterrence strategies.

Strategically, cyberspace is an environment of persistent offense, where attacking is tactically and strategically more advantageous than defending. As Harknett and Goldman (2016) argue, in an offense-persistent environment, defense can achieve tactical and operational success in the short term if it can constantly adjust to the means of attack, but it cannot win strategically. Offense will persist and interactions with the enemy will remain constant.

At the same time, cyber attacks and defense evolve along with digital technology. As the latter becomes increasingly autonomous and smart, leveraging the potential of artificial intelligence (AI) (Yang et al. 2018), so do cyber attacks and defense strategies. Both the public and private sectors are already testing AI systems in autonomous war games (Taddeo and Floridi 2018). The 2016 DARPA Cyber Grand Challenge was a landmark in this respect. The Challenge was the first competition in which AI capabilities for defense were successfully tested and showed to be able to identify and patch their own vulnerabilities, while also countering threats and targeting the vulnerabilities of antagonist systems.

Given the strategic nature of cyberspace and the role of AI in cyber defense, I argued elsewhere (Taddeo 2018) that, to be effective, a theory of cyber deterrence rests on three elements: *target identification*, *retaliation*, and *demonstration* (Fig. 1).

[4] https://www.thecipherbrief.com/column_article/no-thing-cyber-deterrence-please-stop

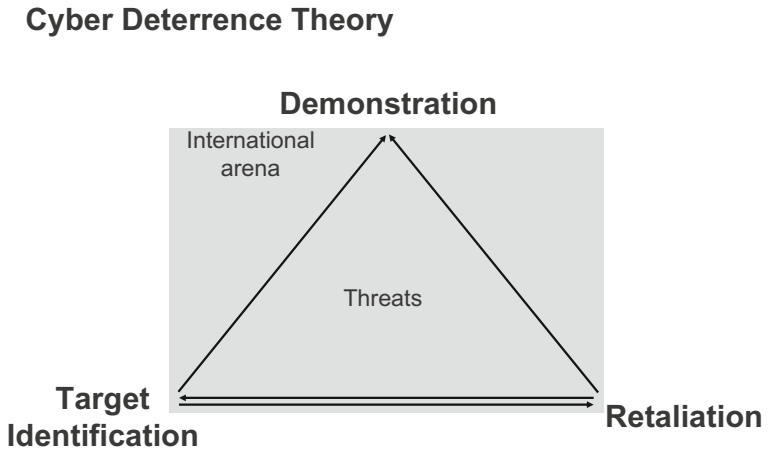

Fig. 1 The three elements of cyber deterrence theory and their dependencies (Taddeo 2018)

According to the model showed in Fig. 1, target identification is essential for deterrence. It allows the defendant to isolate (and counter-attack) enemy systems independently from the identification of the actors behind them, thus side-stepping the attribution problem, while identifying a justifiable target for retaliation. Identifying the attacking system and retaliate is feasible task, one which AI systems for defense can already achieve. Deterrence in cyberspace works by *demonstrating* the defendant's capability to retaliate an occurring attack and harming the opponent's system. While it may not deter an *incoming* cyber attack, retaliation will deter the *next* rounds of attacks coming from the same opponent. This is because, given the offense-persistent nature of cyberspace, the mere threat of retaliation will not be sufficient, at first, to change the opponent's intentions to attacks. The chances of success and the likelihood that the attack will remain unattributed remain too high for threats, albeit credible, to be effective. Thus, to be successful, cyber deterrence need to shift from *threatening* to *prevailing*. By deploying this strategy, States will be able to build a reputation on the basis of their capability and commitment to retaliate, which will lead, over time, to stronger cyber deterrence postures.

While this model would enable deterrence of cyber attacks, by itself it is insufficient to ensure stability of cyberspace. This is true especially when considering how the rising distribution and automation, multiple interactions, and fast-pace performance of cyber attacks make control progressively less effective, while increasing the risks for unforeseen consequences, proportionality breaches, and escalation of responses. An international regime of norms regulating state behaviour in cyberspace is necessary to complement cyber deterrence strategies and foster stability (Taddeo and Floridi 2018).

This is why the 2017 failure of the UN Governmental Group of Experts on "Developments in the Field of Information and Telecommunications in the Context of International Security" (GGE)[5] to provide recommendations on State conduct in cyberspace is problematic (Taddeo 2017a). Over the past 20 years, the UN GGE, the

[5] https://www.justsecurity.org/42768/international-cyber-law-politicized-gges-failure-advance-cyber-norms/

Organization for Cyber Security and Co-operation in Europe (OSCE), and the ASEAN Regional Forum (ARF), and several national governments (G7 and G20) have been building consensus to define such a regime of norms. The time has come now to build on these initiatives and define binding norms for state actors in cyberspace.

These norms will have to be enforced by an independent authority able to exert coercive power and impose sanctions. This authority cannot (and should not) be the result of a multi-stakeholder or a neutral, private-led initiative, as suggested for example by the proposal for a Digital Geneva Convention.[6] This would impose too heavy civil responsibilities on the private sector and create an authority too weak to face the political pressure resulting from ensuring State compliance to the regime of norms.

Enforcing this regime requires an authority able to (i) ensure States compliance with the norms at an international level, (ii) run investigations into suspected State-run (or State-sponsored) cyber operations to define attribution, (iii) expose breaches of the norms, and (iv) impose adequate sanctions and punishments. These requirements define a political mandate for an authority that will have a deep impact on international relations and geo-political equilibriums.

Points (i)–(iv) resonate with Article 26 of the UN Charter, which defines the mission of the Security Council:

> […] to promote the establishment and maintenance of international peace and security with the least diversion for armaments of the worlds human and economic resources, the Security Council shall be responsible for formulating, with the assistance of the Military Staff Committee […] plans for the establishment of a system for the regulation of armaments.[7]

Undeniably, the UN Security Council has the necessary resources, the political, and coercive power to achieve (i)–(iv). The time has come to embrace this power to consolidate and enforce an international regime of norms to regulate state behaviour in cyberspace. Problems, mistakes, and even failures are to be expected, but they must not hinder the process.

This special issue has the goal of landscaping the debate on cyber deterrence and its role in fostering cyber stability. For this reason, it includes contributions focusing on strategies for cyber deterrence as well as articles addressing the ethical and regulatory aspects of state behaviour in cyberspace. More in detail, the first two articles focus on the strategic aspects of cyber deterrence. "Five Kinds of Cyber Deterrence" (Ryan 2017) sets the tone of the issue by mapping the main approaches to cyber deterrence provided in the extant literature. "The Limits of Deterrence Theory in Cyberspace" (Taddeo 2017b) identifies the limits of deterrence theory in cyberspace and define the conceptual space for a domain-specific theory of cyber deterrence. "Just War, Cyber War, and the Concept of Violence" (Finlay 2018) shifts the focus on normative aspects of cyber conflicts to consider whether cyber threats may justifiably be characterized as a form of "violence." "Warfighting for Cyber Deterrence: a Strategic and Moral Imperative"

[6] https://blogs.microsoft.com/on-the-issues/2017/02/14/need-digital-geneva-convention/
[7] http://www.un.org/en/sections/un-charter/chapter-v/

(Lonsdale 2017) also offers a normative analysis of cyber conflicts and deterrence. "Why the World Needs an International Cyberwar Convention" (Eilstrup-Sangiovanni 2017) draws on existing normative regimes for the regulation on the use of weapons to argue for the feasibility of an international convention on the use of cyber weapons. "Deterrence in Cyberspace: a Silver Bullet or a Sacred Cow?" (Lawson 2017) concludes the special issue by considering the way in which different deterrence strategies could be implemented in cyberspace.

Before leaving the reader to this special issue, I would like to express my sincere gratitude to the authors who contributed to it, as well as to the colleagues with whom I have had the opportunity of discussing several of the topics addressed in this issue, in particular Paul Cornish, Grahm Fairclough, and the members of the Digital Ethics Lab of the University of Oxford. I would also like to thank Luciano Floridi, the editor-in-chief of *Philosphy & Technology*, for his support during the preparation of this issue and throughout the process the led to defining many of the ideas presented in this introduction.

References

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

Chadwick, A., & Howard, P. N. (Eds.). (2009). *Routledge Handbook of Internet Politics. Routledge handbooks*. London: Routledge.

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.

Eilstrup-Sangiovanni, M. (2017). Why the world needs an international cyberwar convention. *Philosophy & Technology*. https://doi.org/10.1007/s13347-017-0271-5.

Finlay, C. J. (2018). Just war, cyber war, and the concept of violence. *Philosophy & Technology*. https://doi.org/10.1007/s13347-017-0299-6.

Floridi, L. (2016). Mature information societies—a matter of expectations. *Philosophy & Technology, 29*(1), 1–4. https://doi.org/10.1007/s13347-016-0214-6.

Floridi, L., & Taddeo, M. (Eds.). (2014). *The ethics of information warfare*, *Law, governance and technology series, volume* (Vol. 14). Heidelberg: Springer.

Freedberg, S. (2014). NATO Hews To Strategic Ambiguity On Cyber Deterrence. 2014.

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

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, T., Xin, Z., Raduege Jr., H., Grigoriev, D., Duggal, P., & Schjølberg, S. (2010). Global Cyber Deterrence Views from China, the U.S., Russia, India, and Norway. EastWest Institute.

Lawson, E. (2017). Deterrence in cyberspace: a silver bullet or a sacred cow? *Philosophy & Technology*. https://doi.org/10.1007/s13347-017-0267-1.

Libicki, M. (2009). *Cyberdeterrence and cyberwar*. The RAND Corporation.

Lonsdale, D. J. (2017). Warfighting for cyber deterrence: a strategic and moral imperative. *Philosophy & Technology*. https://doi.org/10.1007/s13347-017-0252-8.

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.

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

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

Taddeo, M. (2012). Information warfare: a philosophical perspective. *Philosophy and Technology, 25*(1), 105–120.

Taddeo, M. (2014). *Just information warfare* (pp. 1–12). April: Topoi. https://doi.org/10.1007/s11245-014-9245-8.

Taddeo, M. (2016). On the risks of relying on analogies to understand cyber conflicts. *Minds and Machines, 26*(4), 317–321. https://doi.org/10.1007/s11023-016-9408-z.

Taddeo, M. (2017a). Deterrence by norms to stop interstate cyber attacks. *Minds and Machines*. https://doi.org/10.1007/s11023-017-9446-1.

Taddeo, M. (2017b). The limits of deterrence theory in cyberspace. *Philosophy & Technology*. https://doi.org/10.1007/s13347-017-0290-2.

Taddeo, M. (2018). How to deter in cyberspace. *The European Centre of Excellence for Countering Hybrid Threats, 2018*(6), 1–10.

Taddeo, M., & Floridi, L. (2018). Regulate artificial intelligence to avert cyber arms race. *Nature, 556*(7701), 296–298. https://doi.org/10.1038/d41586-018-04602-6.

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

Yang, G.-Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., Jacobstein, N., et al. (2018). The Grand Challenges of *Science Robotics*. *Science Robotics, 3*(14), eaar7650. https://doi.org/10.1126/scirobotics.aar7650.