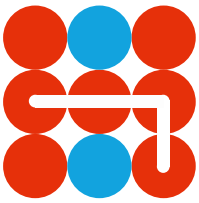
Strategic Analysis June-July 2018

How to Deter in Cyberspace

MARIAROSARIA TADDEO



Hybrid CoE

How to Deter in Cyberspace

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. – writes Dr Mariarosaria Taddeo.

Cyber attacks are escalating in frequency, impact, and sophistication. State actors play an increasingly larger role in this escalating dynamics, as they use cyber attacks both offensively and defensively. For example, North Korea has been linked to WannaCry, and Russia to NotPetya, two major cyber attacks launched in 2017. Russia has also been linked to a series of cyber attacks targeting US critical national infrastructures, which were disclosed in 2018. Concerned by the risks of escalation, international organisations such as NATO, the UN Institute for Disarmament Research (UNIDIR), and national governments in the likes of the UK and the US, have started to consider whether, and how, to deploy deterrence to maintain the stability of cyberspace.

Conventional deterrence theory

Defining cyber deterrence strategies is challenging. **Conventional deterrence theory (henceforth deterrence theory) does not work in cyberspace, as it does not address the global reach, anonymity, distributed and interconnected nature of this domain**. Deterrence theory has three core elements: attribution of attacks; defence and retaliation as types of deterring strategies; and the capability of the defender to signal credible threats (see Figure 1). None of these elements is attainable in cyberspace.

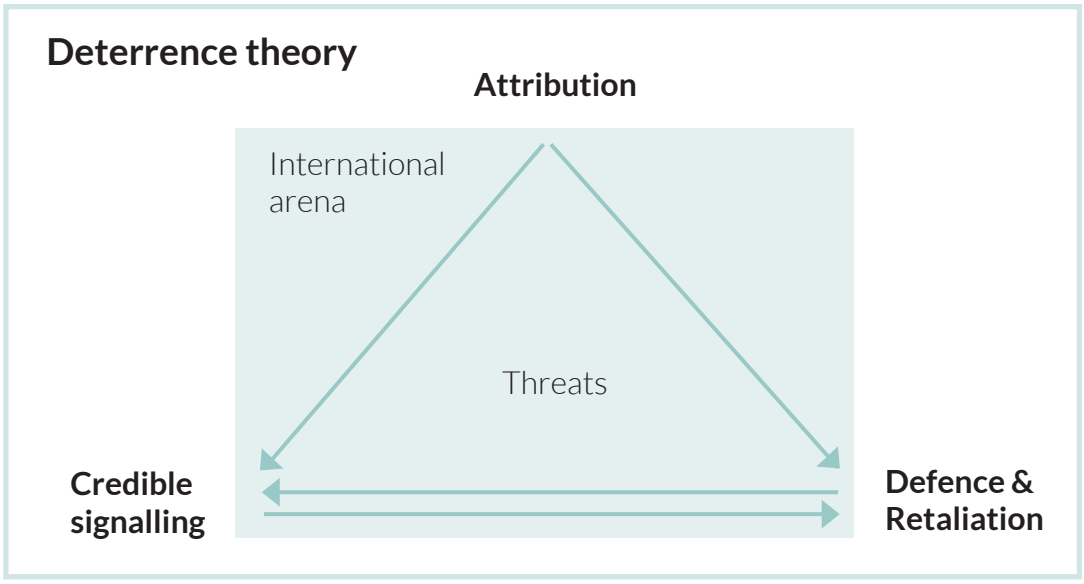


Figure 1. The core elements of deterrence theory and their dependencies. This Figure was first published in M. Taddeo, "The Limits of Deterrence Theory in Cyberspace", *Philosophy & Technology*, 2017.

Attribution is crucial to conventional deterrence

Consider attribution first. Prompt, positive attribution is crucial to deterrence: the less immediate is attribution, the less severe will be the defender's response. The less positive the attribution, the more time will be needed to respond. **In cyberspace, attribution is at best problematic, if not impossible**. Cyber 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. 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 attackers can design and route their operations through third-party machines and countries with the goal of obscuring or misdirecting attribution. **The limits of attribution in cyberspace pose serious obstacles to the deployment of effective deterrence**. Recalling Figure 1, without attribution, defence and retaliation, as well as signalling are left without a target and are undermined by the inability of the defender to identify the attacker.

Signalling credible threats

Signalling is crucial for deterrence; it is the moment at which the defender communicates (threatens) to the attackers that consequences will follow if they decide to attack. In order to be effective, the threats need to be credible. And the credibility of the threats hinges on the state's reputation. In kinetic scenarios, a state's reputation is gained by showcasing military capabilities and by demonstrating a state's ability to resolve conflicts (to deter or defeat the opponent) over time. To some extent, the same holds true in cyberspace, where a state's reputation also refers to its past interactions in this domain, its known cyber capabilities to defend and retaliate, as well as its overall reputation for resolving conflicts. However, signalling credible threats is problematic in cyberspace. This is because **a state's reputation in cyberspace may not necessarily correspond to actual capabilities in this domain, as states are reluctant to circulate information about the attacks that they are subjected to, especially those that they could not avert**. This makes signalling less credible and, thus, more problematic than in other domains of warfare.

Deterrence by defence and by retaliation

Deterrence by defence is guaranteed to be ineffective in cyberspace. Defence in cyberspace is porous; every system has its security vulnerabilities and identifying and exploiting them is simply a matter of time, means, and determination. This makes even the most sophisticated defence mechanisms ephemeral, thereby limiting their potential of defence to deter new attacks. Even when successful, cyber defence does not lead to a strategic advantage, insofar as averting a cyber attack very rarely leads to the ultimate defeat of an adversary.

Unlike deterrence by defence, deterrence by retaliation may be effective in cyberspace. However, this strategy is coupled with a serious risk of escalation. This is because the means to retaliate, namely

cyber weapons, are malleable and difficult to control. **Cyber weapons can be accessed, stored, combined, repurposed, and redeployed much more easily than was ever possible with other kinds of military capability**. This was the case with Stuxnet, for example. 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 US.

We need to evolve our thinking

Clearly, deterrence theory faces severe limitations when applied in cyberspace. But it would be a mistake to conclude that as deterrence theory does not work in cyberspace, then deterrence is unattainable in this domain. As USN Lt. Commander Robert Bebber stated:

"History suggests that applying the wrong operational framework to an emerging strategic environment is a recipe for failure. During 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". [1]

We need to adapt or, better, we need to evolve our way of thinking about cyber conflicts. In turn, this requires **an in-depth understanding of cyberspace, cyber conflicts, their nature, and their dynamics**. This understanding will allow us to forge a new theory of deterrence, one that is able to address the specificities of cyberspace and cyber conflicts. **The alternative – developing cyber deterrence by analogy with conventional deterrence – is a recipe for failure**. It is equivalent to forcing the proverbial square peg into a round hole: we are more likely to smash the toy than to win the game.

In defining a theory for cyber deterrence, it is important to take into account both the nature of cyberspace and the new capabilities that digital technologies bring about. **Cyberspace is an environment of persistent offence, where attacking is tactically and strategically more advantageous than defending**. As Harknett and Goldman (2016) argue, in an offence-persistent environment, defence 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. **Offence will persist and interactions with the enemy will remain constant. This is why inter-state cyber defence has shifted from reactive (defending) towards active (countering) defence strategies**.

At the same time, cyber attacks and defence evolve along with digital technology. As the latter becomes more autonomous and smart, leveraging the potential of artificial intelligence (AI), so do cyber attacks and cyber defence strategies. Both the public and private sectors are already testing AI systems in autonomous war games. The 2016 DARPA Cyber Grand Challenge was a landmark in this respect. The Challenge was the first fully autono-

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

mous competition in which AI capabilities for defence were successfully tested. Seven AI systems, developed by teams from the United States and Switzerland, fought against each other to identify and patch their own vulnerabilities, while probing and exploiting those of other systems. The Challenge showed that AI will have a major impact on the waging of cyber conflicts; it will provide new capabilities for defence and shape new strategies, but also pose new risks. The latter are of particular concern. The autonomy AI systems, their capacity to improve their own strategies and launch increasingly aggressive counter-attacks with each iteration, may lead to breaches of proportionality and escalation of responses, which could, in turn, trigger kinetic conflicts. In this scenario, cyber deterrence is ever more necessary.

Cyber deterrence theory

A theory of cyber deterrence rests on three core elements: target identification, retaliation, and demonstration (Figure 2).

Target identification is essential for deterrence. It allows the defender to isolate (and counter-attack) enemy systems independently from the identification of the actors behind them, thereby side-stepping the attribution problem, while identifying a justifiable target for retaliation. Identifying the attacking system and retaliating is a feasible task, and one which AI systems for defence can already achieve. As shown in Figure 2, cyber deterrence does not encompass defence among its possible strategies. This is due to the offence-persistent nature of cyberspace, which makes retaliation more effective than defence both tactically and strategically.

Cyber deterrence uses target identification and retaliation for demonstrative purposes. **According to this theory, deterrence in cyberspace works if it can demonstrate the defender's capability**

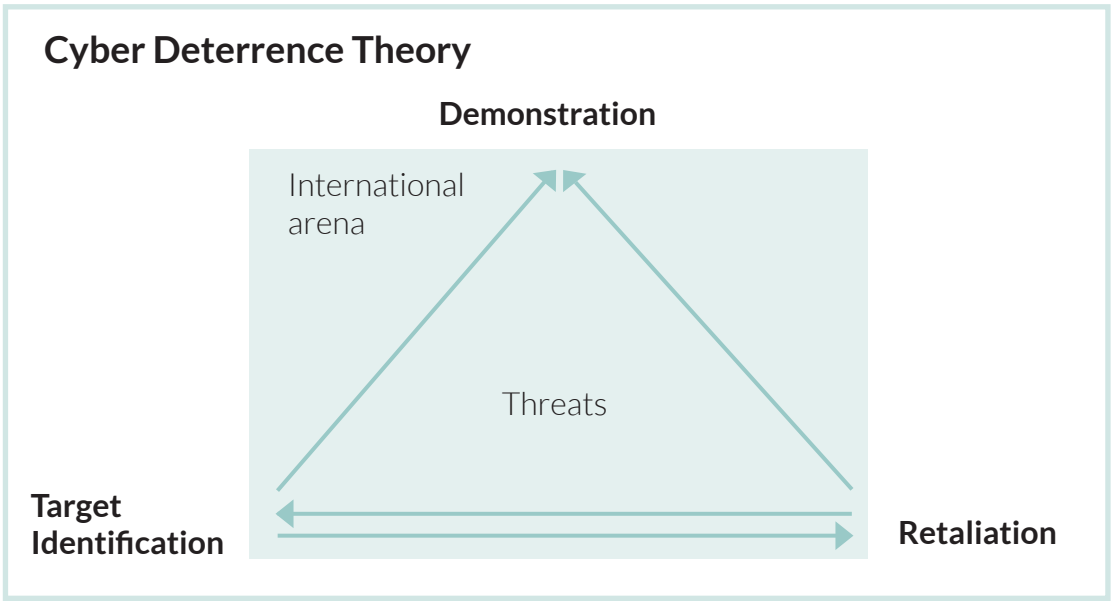


Figure 2. The three elements of Cyber Deterrence Theory and their dependencies

to retaliate against a current attack by harming the source system. While not being able to deter an incoming cyber attack, retaliation will deter the next round of attacks coming from the same opponent. This is because the mere threat of retaliation will not be sufficient to change the opponent's intention to attack. The chances of success and the likelihood that the attack will remain unattributed remain too high for any proportionate threat to be effective. **Thus, to be successful, cyber deterrence needs to shift from threatening to prevailing**.

International regime of norms

Cyber deterrence alone is not a panacea: it is necessary but insufficient to ensure the stability of cyberspace. This is true especially when considering how the rising distribution and automation, multiple interactions, and fast-paced performance of cyber attacks make control progressively less effective, while increasing the risks of 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 maintain stability**. Three steps are crucial to this end:

- **Define 'red lines'** distinguishing legitimate and illegitimate targets and definitions of proportionate responses for cyber defence strategies.
- **Build alliances** by mandating 'sparring' exercises between allies to test AI-based defence capabilities and the disclosure of fatal vulnerabilities of key systems and crucial infrastructures among allies.
- **Monitor and enforce rules at an international level** by defining procedures to audit and oversee AI-based state cyber defence operations, alerting and redressing mechanisms to address mistakes and unintended consequences. A third-party authority with teeth, such as the UN Security Council, should rule on whether red lines, proportionality, responsible deployment or disclosure norms have been breached.

Deterrence in cyberspace is possible. But it requires an effort to develop a new domain-specific, conceptual, normative, and strategic framework. Analogies with deterrence theory and existing normative frameworks should be abandoned altogether, as they are misleading and detrimental to any attempt to develop an innovative and in-depth understanding of cyber deterrence and to ensure stability of cyber space. **The effort is be complex, but also feasible. The time has come to join forces and begin working to achieve deterrence and regulation of state behaviour in cyberspace. These are both key elements in avoiding escalation and in ensuring stability**.

Author

Dr Mariarosaria Taddeo is Research Fellow at the Oxford Internet Institute, University of Oxford, where she is the Deputy Director of the Digital Ethics Lab, and Faculty Fellow at the Alan Turing Institute. Her recent work focuses mainly on the ethical analysis of cyber security practices, interstate cyber conflicts, and the ethics of AI. Her area of expertise is Philosophy and Ethics of Information, although she has worked on issues concerning Epistemology, Logic, and Philosophy of AI. Dr Taddeo has been awarded the Simon Award for Outstanding Research in Computing and Philosophy. She also received the World Technology Award for Ethics, acknowledging the originality of her research on the ethics of cyber conflicts, and the social impact of the work that she developed in this area. Since 2016, Dr Taddeo serves as editor-in-chief of *Minds & Machines* (SpringerNature) and as editor-in-chief of *Philosophical Studies Series* (SpringerNature). Her most recent research has been published in *Science Robotics and Nature*.

Further reading:

Arquilla, John. (2013). "Twenty Years of Cyberwar." *Journal of Military Ethics* 12 (1): 80–87. https://doi.org/10.1080/15027570.2013.782632

Floridi, L., and M. Taddeo, eds. (2014). *The Ethics of Information Warfare*. New York: Springer.

Freedberg, Sydney. (2014). "NATO Hews To Strategic Ambiguity On Cyber Deterrence." Breaking Defense, November. https://breakingdefense.com/2014/11/natos-hews-to-strategic-ambiguity-on-cyber-deterrence/

Freedman, Lawrence. (2004). Deterrence. Cambridge, UK; Malden, MA: Polity Press.

Harknett, Richard, J., and Emily O. Goldman (2016). "The Search for Cyber Fundamental." *Journal of Information Warfare* 15 (2): 81–88.

International Security Advisory Board. (2014). "A Framework for International Cyber Stability." United States Department of State. http://goo.gl/azdM0B

Libicki, Martin. (2009). *Cyberdeterrence and Cyberwar*. The RAND Corporation.

Nye, Joseph S. (2011). "Nuclear Lessons for Cyber Security?". *Strategic Studies Quarterly 5* (4): 11–38.

Taddeo, Mariarosaria. (2014). "Just Information Warfare." *Topoi*, April, 1–12. https://doi.org/10.1007/s11245-014-9245-8

Taddeo, Mariarosaria. (2016). "On the Risks of Relying on Analogies to Understand Cyber Conflicts." *Minds and Machines* 26 (4): 317–21. https://doi.org/10.1007/s11023-016-9408-z

Taddeo, Mariarosaria. (2017a). "Deterrence by Norms to Stop Interstate Cyber Attacks." *Minds and Machines*, September. https://doi.org/10.1007/s11023-017-9446-1

Taddeo, Mariarosaria. (2017b). "The Limits of Deterrence Theory in Cyberspace." *Philosophy & Technology*, October. https://doi.org/10.1007/s13347-017-0290-2

Taddeo, Mariarosaria, and Luciano Floridi. (2018). "Regulate Artificial Intelligence to Avert Cyber Arms Race." *Nature 556* (7701): 296–98. https://doi.org/10.1038/d41586-018-04602-6

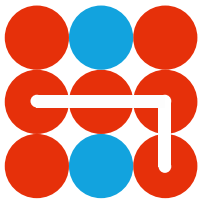
Taddeo, Mariarosaria, and Ludovica Glorioso, eds. (2016). *Ethics and Policies for Cyber Operations*. Philosophical Studies, Springer.

UN Institute for Disarmament Research. (2014). "Cyber Stability Seminar 2014: Preventing Cyber Conflict."

The European Centre of Excellence for Countering Hybrid Threats
tel. +358 400 253800 www.hybridcoe.fi

ISBN: 978-952-7282-09-0

Hybrid CoE is an international hub for practitioners and experts, building member states' and institutions' capabilities and enhancing EU-NATO cooperation in countering hybrid threats located in Helsinki, Finland



Hybrid CoE