Regulate Artificial Intelligence To Avert Cyber Arms Race

Define an international doctrine for cyberspace skirmishes before they escalate into conventional warfare, urge **Mariarosaria Taddeo** and **Luciano Floridi**.

Mariarosaria Taddeo[1,2,3*] and Luciano Floridi[1,2,3]

[1]Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford, OX1 3JS, United Kingdom;

[2]The Alan Turing Institute, 96 Euston Road, London, NW1 2DB, United Kingdom;

[3]Department of Computer Science, University of Oxford, Oxford OX1 3QD, United Kingdom.

[*]Corresponding author: mariarosaria.taddeo@oii.ox.ac.uk

Cyber-attacks are becoming more frequent, sophisticated, and destructive. Each day the US suffers more than 4,000 ransomware ttacks that encrypt computer files until the owner pays to release them [1]. This is triple the number in 2015. In May 2017, when WannaCry virus crippled hundreds of IT systems across the UK's National Health Service, 20,000 appointments were cancelled. A month later, the NotPetya ransomware cost pharmaceutical giant Merck, shipping firm Maersk, and logistics company FedEx around $300 million each in immediate damages. Global damages from cyber-attacks totalled $5 billion in 2017 and may reach $6 trillion a year by 2021. States are partly behind this rise. They use cyber-attacks both offensively and defensively. For example, North Korea has been linked to WannaCry, Russia to NotPetya.

As the threats escalate, so do defence strategies. Since 2012, the US has employed 'active' cyber defence strategies that enable computer experts to neutralise or distract viruses with decoy targets, and to break back into a hacker's system to delete data or to destroy it completely. In 2016, the UK announced a £1.9bn investment and a 5-year plan to combat cyber threats. NATO also began drafting principles for such operations, to be agreed by 2019. The US and the UK are leading this initiative, with Denmark, Germany, the Netherlands, Norway, and Spain also involved [2].

Artificial intelligence (AI) is poised to revolutionize this activity. Attacks and responses will become faster, more precise, and more disruptive. Threats will be dealt with in hours, not days or weeks. AI solutions are being rolled out that can verify code and identify bugs and vulnerabilities, making the digital systems they protect more robust. For example, in April 2017 software firm DarkTrace launched Antigena, which uses machine learning to spot abnormal behaviour on an IT network, shut off communications to that part of the system, and issue an alert. Implementations of AI in cyber security

will become increasingly pervasive. According to some estimates, the value of AI in cyber security will reach $18.2 billion by 2023 (Figure 1).

Experts welcome the potential of AI to improve system verification and robustness [6]. However, this potential is crippled by two risks: the escalation of responses and the lack of control over snowballing effects that AI-led counter-attacks may have. Thus, it is crucial that experts ensure transparency about problems, limitations, and possible shortcomings of unregulated uses of AI for defence, while also working with policy-makers and end users to design adequate testing and oversight mechanisms for this technology.

By the end of this decade, many states plan to deploy AI for national cyber defence [4,5]. AI makes deterrence possible [3] because algorithms can identify the source and neutralise it without having to identify the actor behind it, so that attacks can be punished. Currently, countries hesitate to push back because they are unsure who is responsible, given that campaigns may be waged through third-party computers and often use common software.

The risk is a cyber arms-race [6]. As states use increasingly aggressive AI-driven strategies, opponents will respond ever more fiercely. Such a vicious cycle might lead ultimately to a physical attack.

States need regulations that establish legitimate uses of AI for national defence, shape the process leading to the deployment of AI, such as testing and auditing mechanisms, and define the chain of responsibilities [7]. Cyberspace is a domain of warfare and AI is a new defence capability, so regulations are necessary for state use of AI, much like there exist regulations for the other domains of warfare – air, see, land, and space – and the corresponding military capabilities. What is long overdue are criteria for determining proportional responses; clear thresholds, so called ‘red lines’, for distinguishing legal and illegal cyber-attacks; and appropriate sanctions for illegal acts [8]. In each case, unilateral approaches will be ineffective. Rather, it is crucial to define an international doctrine for state action in cyberspace. Alarmingly, international efforts to regulate cyber conflicts have stalled. We call on regional forums, like NATO and EU, to revive efforts and prepare the ground for an UN-led initiative.

No rules

Right now the UN process is in deadlock. In 2004, the United Nations set up a Group of Governmental Experts (UN GGE) to agree on a set of voluntary rules for how states should behave in cyberspace. Their 5th meeting in 2017 ended in a stand-off. The Group could not reach consensus on whether existing international laws on self-defence, humanitarian laws, and state responsibility should apply in cyberspace. The US argued cyber-defence regulations should build on these foundations. Other nations, including Cuba, Russia, and China**,** firmly disagreed, arguing that this would ‘militarize’ cyberspace, sending the wrong message about peaceful conflict resolution. The Group failed to deliver its report. It is unclear whether it will meet again, or what will happen next.

International dialogue and action must resume. NATO could pave the way through its forthcoming guidelines. Although it is currently unclear what their scope will be.

Meanwhile, research on AI for cyber defence is progressing fast. The US is in the lead, technologically. It aims to incorporate AI into its cyber-defence systems by 2019 [4]. The Department of Defense (DoD) has earmarked $150 million for research. The US Defense Advanced Research Projects Agency (DARPA) is developing techniques to do this. Steps have already been taken. In its 2016 Cyber Grand Challenge competition, seven AI systems, developed by academics and private companies from Greece, Russia, Switzerland, and the US played against each other, identifying and targeting opponents’ weaknesses, while finding and patching their own.

The DoD will issue the first US report on AI strategies for national defence in May. There is no indication of what its approach will be. Previous documents, like the 2015 US Cyber Strategy or the 2016 National Cyber Incident Response Plan, did not cover autonomous systems, machine learning, or AI. The 2012 DoD Directive on ‘Autonomy in Weapon Systems’ [9] focused on internal procedures for deploying AI, but was silent on when the US would do so in the international arena.

AI is a priority for China, which aims to become a world leader in machine learning technologies. In July 2017, the Chinese government issued its AI Development Plan [5]. Military implementation of AI, on the battlefield as well as in cyberspace, is a crucial part of the strategy. But it is unclear to what degree China plans to deploy AI actively in cyber defence.

Russia has not released any public documents about its strategies for AI in defence. However, in a video message released in 2017, President Putin expressly referred to AI and stated “whoever becomes the leader in this sphere will become the ruler of the world”. Experts agree that Russia is focusing on developing AI-enhanced tools for its conventional forces. However, since 2014, the Russian National Defense Control Center (NDCC) has also used machine learning algorithms to detect online threats. Allegedly, Russia has pioneered the use of AI to spread disinformation and intervene in the public debate of other nations, like in the case of the US Presidential election and the UK European Union membership referendum, in 2016. Although these operations are not part of national defence strategies, they speak to the level of AI capabilities developed by Russia over the past two years.

North Korea has a history of cyberspace aggression. WannaCry in 2016 and the cyber-attack against Sony Picture in 2014 are two good example of North Korea aggressive attitude in cyberspace. The country lacks technical expertise in AI but is likely to want to catch up with its adversaries.

The EU is stepping-up too. In 2017, it reassessed cyber security and defence policies and launched the European Centre of Excellence for Countering Hybrid Threats. EU directives present the most comprehensive regulatory framework for state conduct in cyberspace to date. Yet they do not go far enough. The EU treats cyber defence as a case of cyber security, to be improved passively by making member states' information systems more resilient. It disregards active uses of cyber defence and does not include AI. This is a missed opportunity. The EU could have begun defining 'red lines' and proportionate responses. For example, the EU Directive on Security of Network and Information Systems [10] provides criteria for identifying critical national infrastructures, such as health systems or key energy and water supplies that should be protected. The same criteria could be used to define illegitimate targets of state-sponsored cyber-attacks.

Regional forums, like NATO and the EU, must take the following three steps to avoid serious imminent attacks on state infrastructures and to maintain international stability.

Three steps

Define legal boundaries. The international community needs to agree urgently upon 'red lines' distinguishing legitimate and illegitimate targets and definitions of proportionate responses for cyber defence strategies. International consensus at the UN level will ultimately be required. Until then guidelines from regional multilateral bodies, like NATO and the EU, must cover these issues and lead by example.

Test strategies with allies. 'Sparring' exercises should be organised between friendly countries to test AI-based defence tactics. These tests should be mandatory before any system is deployed. They could be in the form of DARPA's Grand Challenge or the cyber simulation exercises routinely run by NATO and the EU. Because AI learns by experience, these matches will improve the strategies of the alliance, while finding and healing weaknesses. Fatal vulnerabilities of key systems and crucial infrastructures should be shared with allies; policy frameworks should demand disclosure. Agreements and regulations with similar sharing and disclosure requirements include the EU eIDAS (electronic IDentification, Authentication and trust Services) Regulation and the IPA (Industry Partnership Agreement) within NATO.

Monitor and enforce rules. The international community needs to agree on how to audit and oversee AI-based state cyber defence operations. Alert and remedy mechanisms are needed 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. As in the case with other international norms, economic or political sanctions should be imposed on states violating rules. NATO and the EU should enforce the norms within their remits.

The solution is difficult, but it is clear. There is no time to waste.

References

[1] Federal Bureau of Investigation, “Ransomware Prevention and Response for CISOs,” 2017.

[2] “NATO mulls ‘offensive defence’ with cyber warfare rules,” *Reuters*, 30-Nov-2017.

[3] M. Taddeo, “The Limits of Deterrence Theory in Cyberspace,” *Philos. Technol.*, Oct. 2017.

[4] Defence Science Board, “Summer Study on Autonomy,” *US Dep. Def.*, Jun. 2016.

[5] Chinese State Council, “New Generation Artificial Intelligence Development Plan,” Chinese State Council, Jul. 2017.

[6] G.-Z. Yang *et al.*, “The grand challenges of *Science Robotics*,” *Sci. Robot.*, vol. 3, no. 14, p. eaar7650, Jan. 2018.

[7] L. Floridi, “Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions,” *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, vol. 374, no. 2083, p. 20160112, Dec. 2016.

[8] M. Taddeo, “Deterrence by Norms to Stop Interstate Cyber Attacks,” *Minds Mach.*, Sep. 2017.

[9] Department of Defence, “Directive on Autonomy in Weapon Systems,” Washington, D.C, USA, 2012.

[10] The European Parliament and the Council of The European Union, “Directive on security of network and information systems.” 07-Jun-2016.