Minds and Machines (2019) 29:635–645
https://doi.org/10.1007/s11023-019-09508-4

COMMENTARY

# Ethics of AI and Cybersecurity When Sovereignty is at Stake

**Paul Timmers**[1]

Published online: 11 October 2019


**Abstract**
Sovereignty and strategic autonomy are felt to be at risk today, being threatened by the forces of rising international tensions, disruptive digital transformations and explosive growth of cybersecurity incidents. The combination of AI and cybersecurity is at the sharp edge of this development and raises many ethical questions and dilemmas. In this commentary, I analyse how we can understand the ethics of AI and cybersecurity in relation to sovereignty and strategic autonomy. The analysis is followed by policy recommendations, some of which may appear to be controversial, such as the strategic use of ethics. I conclude with a reflection on underlying concepts as an invitation for further research. The goal is to inspire policy-makers, academics and business strategists in their work, and to be an input for public debate.

**Keywords** Cybersecurity · Ethics · Artificial intelligence · Sovereignty · Strategic autonomy

## 1 Sovereignty and Strategic Autonomy in the Digital Age

Over the last few years strategic autonomy and sovereignty have become top political priorities. Government leaders feel that national sovereignty is under threat. The reason is a confluence of pervasive, transformative and even disruptive digital technologies, explosive growth of cyber incidents, and rising international tensions between the US and EU on one side and China and Russia at the other side, as well as transatlantic tensions.

There is no doubt that these threats put sovereignty at stake. Kello (2017) argues that 'cyber' creates a 'sovereignty gap'. Both state and non-state actors are exploiting cybersecurity means. Kello observes a combination of persistent disruption ('unpeace'), rogue state actors that misuse cyber technologies, and cyber-enabled exercise of influence by non-state actors, from state-proxies (Maurer 2018) to

✉ Paul Timmers
paul.timmers@politics.ox.ac.uk

[1] University of Oxford, Oxford, UK

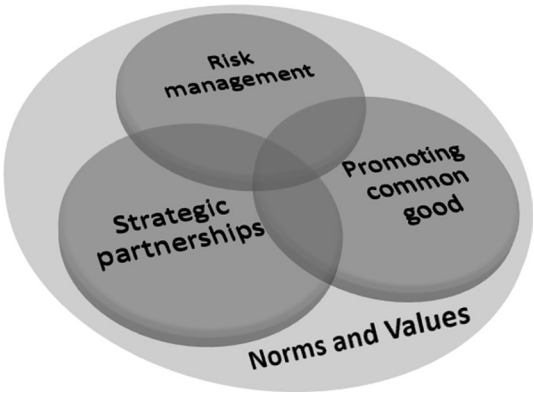


**Fig. 1** Approaches to address strategic autonomy in relation to cybersecurity

terrorists to global platforms, that systemically alter the balance of power in the traditional state-based (Westphalian) system of international relations.

Policy-makers and politicians tend to see strategic autonomy as a means to an end, namely sovereignty. They often join up the 'sovereignty' or 'strategic autonomy' with a term that stands for a critical asset: data sovereignty, digital sovereignty, technological sovereignty, strategic autonomy in defence and military, financial strategic autonomy, and so on.

I define strategic autonomy as "the ability, in terms of capacity and capabilities, to decide and act upon essential aspects of one's longer-term future in the economy, society and their institutions" (Timmers 2019a). Contrary to the past, when strategic autonomy was a term used mostly by France in the military and defence domain and by India to emphasize its foreign policy independence, strategic autonomy nowadays concerns much of economy and society, as well as democracy (think of fake news during elections).

States generally follow three approaches to deal with the challenge of strategic autonomy in the digital age (see Fig. 1). These are: (1) risk management, i.e. keeping the risks to sovereignty manageable as much as possible, which emphasises (cyber-)resilience, (2) strategic partnerships of like-minded states and possibly including private actors to have control on the most critical technologies and systems, and (3) promoting global common goods, to develop and protect certain critical digital assets as a common global interest. A state can pursue one or several of these approaches at the same time.

A fourth approach, i.e. going it completely alone, is at most feasible the US or the People's Republic of China. This approach appears to become increasingly popular in these countries despite dire consequences for global trade, as it is inefficient and requires decoupling of globally interwoven supply chains.

Let's analyse each of the three approaches, from the perspective of sovereignty being as stake and focusing on the ethical aspects of the use of AI.

## 2 Ethical Challenges for AI and Cybersecurity in a Risk Management Approach

A risk management approach seeks to strengthen each of the steps "identify, protect, detect, defend, recover" in relation to risks, notably of critical infrastructures such as electricity, water, health, cloud services, etc. The approach involves large scale sensoring/monitoring of complex assets; big data-based threat detection and analysis; real-time response interpreting business, legal, and ethical rules; and managed infrastructure recovery.

In each of these, AI is considered an essential aid and is already becoming big business. Only with AI is it possible to quickly sift through billions of sensor data points so that the responsible CERT[1] can focus on a handful of noteworthy situations only. The New York Stock Exchange reportedly is attacked half a trillion times a day, with 30–40 attacks of consequence.[2] Providers of AI-based cyber-resilience solutions are already multi-billion-dollar companies.

What are the ethical challenges in cybersecurity risk management, notably when making use of AI? Extensive monitoring and pervasive risk-prevention with the help of AI can be highly intrusive and coercive for people, whether employees or citizens. AI can also be so powerful that people feel that their sense of being in control is taken away. They may get a false sense of security too. Deep-learning AI is, as of today, not transparent in how it reaches a decision from so many data points, yet an operator may blindly trust that decision. AI also can incite freeriding as it is tempting to offload responsibility onto 'the system'.

Risk management is also an approach that accepts a residual risk. Financially this may be offset by cyber insurance, but a political and sovereignty question is how many lost lives are acceptable until internal legitimacy of the state and thereby sovereignty is really at risk (the 2017 Wannacry attack that affected many UK hospitals may have led to the loss of lives). This political question becomes even more sensitive when it is an AI system that autonomously invokes a cyber-defensive strategy, such as shutting down part of the electricity grid which implies a choice which people to put at risk or not.

Technical experts also argue that systems are so complex that they can never be fully protected. The fear is that risk management may not detect the presence of a 'kill switch' in a system which could be activated in international conflict or by accident and shutdown a critical infrastructure such as tele-communications (such arguments have been put forward in the 5G/Huawei debate). Alternatively, the fear is for systematic below-the-radar leakage of intellectual property, which eroding long-term national competitiveness. The role of malicious AI would be to keep such a kill-switch or systematic leakage hidden.

[1] CERT = Computer Emergency Response Team, also called CSIRT: Computer Security Incident Response Team.

[2] Hacking Our Security: Digital Resilience for the Next Cyber Threat, interview with Ray Rothrock (RedSeal), Nov 20, 2018, https://www.computerhistory.org/atchm/hacking-our-security-digital-resilience-for-the-next-cyber-threat/.


**Fig. 2** International norms/National cyber strategies

We are therefore confronted with a plethora of ethical issues when combining AI and cybersecurity in a risk management approach to strategic autonomy. They include erosion of individual autonomy, unfair allocation of liability, the fallacy of human in the loop, the contestable ethics of mass surveillance and of trading off individual casualties versus collective protection.

Internationally, risk management is a fruitful and even the main area for developing norms and values of state behaviour. The UN Governmental Group of Experts has developed norms and principles for stability and restraint (mutual responsiveness), open information and transparency, and compatible governance (Heinl 2019 and Timmers 2019b), see Fig. 2.

Likewise, private–public initiatives such as the Paris Call for Trust and Security in Cyberspace have put forward such norms. Norms become concrete through Confidence Building Measures (CBMs).

An example of a restraint norm is 'do not harm', i.e. a commitment to not attack each other's critical infrastructures. Transparency confidence building measures include information exchange on cyber threats and joint cyber exercises. A much more ambitious transparency CBM would be mutual software code inspection by an independent party. An example of compatible governance CBM could be for governments to agree on consultation on say, on cyber resilience in the health sector, with WHO, global industry and civil society and to 'compare notes'.

Clearly, a restraint norm like 'do not harm' has an ethical basis. Likewise, transparency includes commitment to the ethics of 'do not deceive', and compatible governance includes a commitment 'to be fair, equitable and inclusive'.

In practice, it is hard to successfully implement cyber-CBMs. So far, they only seem to work where strategic autonomy is least at stake such as the Global Alliance against Child Abuse and assistance in awareness raising, training of law enforcement, and national strategy development. In most critical infrastructures, global

collaboration on information exchange and CERT-like capacity building is still a dream. Nevertheless, a strong case could be made to at least collaborate, as suggested, on resilience for the most 'civilian' of critical infrastructures, namely health.

CBMs also only work where there is a credible guarantee of effectiveness, which again has ethical aspects. Let's consider here code inspection. Huawei's code inspection approach in the UK (involving GCHQ) is claimed to be flawed amongst others as it does not include one influential party, namely the Chinese government, who supposedly might make Huawei to implant vulnerabilities in its products (Drew and Parton 2019).

There is another effectiveness challenge, specific to software, namely the frequent updating that far outpaces what manual inspection can keep up with. How do you know that the vendor is honest and diligent in doing these updates? Nevertheless, there may be some light thanks to blockchain-based 'locking' of software versions and AI-based software inspection to ensure such binary equivalence.

Effectiveness of code control is further challenged. Software algorithms are often proprietary, having a high intellectual property value and are therefore not independently inspectable. Moreover, neural network-based AI cannot explain yet how it gets to a decision and is vulnerable to data bias and data poisoning. Effective transparency would then also have to address data input, storage and transmission. Is it ethical to accept such an effectiveness gap?

In short, a risk management approach to strategic autonomy—even if it is the most followed approach today—leaves us with a host of uncomfortable questions on ethics, not exclusive to the application of AI, but certainly exacerbated by AI.

## 3 Ethical Challenges for AI and Cybersecurity in a Strategic Partnership Approach

Let's recall that the strategic partnership means working with sufficiently trusted partners only and in areas that are the most critical; traditionally, that would mostly comprise military systems. Today, however, as strategic autonomy concerns much of economy, society and democracy, the questions are: what is so critical in economic, societal and democratic systems that it should be developed with or supplied by trusted partners only and who are these trusted partners? Recently, Germany proposed a Europeans-only cloud, GAIA-X. Former French Minister Gérard Collomb talked of Franco-European strategic autonomy.

In strategic partnership thinking, AI and cybersecurity takes three forms: (1) AI as a component for the security and safety of critical infrastructures—think of telecoms, smart grids, industry 4.0, or democratic and judicial processes (2) securing the AI that is enabling smart critical facilities such as to prevent hacking of algorithms that control self-driving cars, and (3) weaponized AI, that is AI in cyber- or cyber-kinetic weapons.

Strategic partnerships are with like-minded parties. Such 'like-mindedness' extends to ethics in relation to these first two forms of AI and cybersecurity. Recently the European Commission's high-level group on AI and ethics put forward AI and ethics guidelines (European Commission 2019). Adherence to such guidelines will

become part of the political debate on strategic partnerships. This is the kind of discussion that is familiar from personal data protection and the related EU law, the General Data Protection Regulation (GDPR). Where Europeans stress personal data protection as a human right by law (the GDPR is based on the corresponding Art 16 of the Treaty on the Functioning of the EU), other states consider the GDPR a tool to erect a trade barrier and accuse the EU of using the GDPR for strategic trade geopolitics.

Likewise, the EU must anticipate that its AI and ethics guidelines and possible future legislation on AI will not be seen by everyone as an expression of human rights but rather as a tool of trade politics, as strategic use of ethics. Indeed, we need a debate on perception and reality of 'strategic ethics', even if that may be controversial.

Pursuing a strategic partnership approach to strategic autonomy is clearly a highly political matter. Actors must also be able to steer the direction of partnerships and find common ground, like-mindedness. Doing so they must be able to embed or adapt their own values, in this case to ethics and AI (Taddeo and Floridi 2018) and thereby accept a degree of shared or pooled sovereignty.

AI and cybersecurity for (potentially) offensive purposes ranges from the singular kill-switches (in the past also called 'logic bomb') to AI-based cyber-attack or counter-attack software such as for cyber-deterrence (Taddeo 2018). Such cyber-AI can be combined with physical, kinetic weapons. The spectrum of weaponized AI includes Lethal Autonomous Weapons (LAWS). Many of the ethical issues related to smart weapons are discussed in by Brundage (2018).

In conclusion, the focus in the second approach, strategic partnerships, is on the one hand strategic use of ethics or 'strategic ethics', and on the other hand the ethics of AI-enabled cyber- and cyber-kinetic weapons. While there is much attention for the latter, the former needs a more serious debate to determine the value and viability of a strategic partnership relative to a risk management or global common good approach.

## 4 Ethical Challenges for AI and Cybersecurity in a Global Common Good Approach

Pursuing a global common good approach is not unfamiliar terrain internationally. In the 1980s, a dramatic global challenge was identified: the growing hole in the ozone layer. In response, scientists, policymakers and industry joined forces to reduce the emission of CFCs, the chemicals that were breaking down ozone. Within 2 years, the Montreal Protocol was signed, CFCs were banned and—though it lasted many years—the ozone layer has started to recover. This is a major success in protecting a global common good. Interestingly, in the Montreal approach sovereignty concerns were held in check by the common concern and precautionary principles (Green 2009).

The question is how much of cyberspace can be positioned as a global common good and what is the way to treat it as such, that is, the appropriate governance. The original internet was indeed a "free, open and global internet" and some of

its—perhaps idealistic—creators wanted it to be available as a common good for all of humanity (Barlow 1996). Nevertheless, it came with design flaws that at least partly are the cause of today's cybersecurity threats: it was lacking security-by-design and privacy-by-design. In correcting such flaws AI will play a major role, next to technologies such as blockchain and encryption.

The underlying premise for considering (part of) cyberspace as a global common good is that there is a 'common ethics of global cyberspace'. Technically implementing such global ethics could then imply security-by-design, privacy-by-design, autonomy-by-design and inclusivity.

One might argue, that this is an illustration of the shift from 'code is law' to 'law is code'. That is in the early 2000s Lessig (2000) and Lessig (2006) showed that the technical architecture of the internet conditioned the legislative rules: 'code is law'. Instead, today we rather see 'law is code'. That is, the rules we want to have—as international community or as individual states and whether ethical guidance or hard law—conditions the technologies we accept and allow in the market such as through certification. For AI this would imply, amongst others, a strong emphasis on open source and distributed control. A practical yet highly relevant case to consider is the protection of the public core of the Internet, the domain name system, and to declare that as a global common good (Broeders 2017).

The global common good approach is, by definition, not state-centric. It transcends states and thereby softens the contest on sovereignty. This may seem idealistic and unrealistic, but it is not. It allows states to concentrate their scarce resources for the defence of sovereignty on other matters that cannot be in the global common good, such as military, or justice or even education systems. In that sense such an approach is, to speak with Alexis de Tocqueville, wise from the perspective of "self-interest properly informed" or "enlightened self-interest".

A weakness of the global common good approach for cyberspace is that the required international governance is not in place and existing international internet governance is not well-equipped. Any international governance is also potentially vulnerable for state-centric behaviour through covert capture. Moreover, it is not clear how such international governance can look like.

Nevertheless, hints have been given by (Cowhey and Aronson 2017) and there is a basis to build upon: the UN initiatives (in AI and in cybersecurity), the internet community and related initiatives such as ICANN as well as the Web Foundation's #ForTheWeb movement and the SOLID technical implementation of data protection, and global business initiatives such as oneM2 M for Internet of Things.

Interestingly, most of the UN cybersecurity norms, principles, rules and CBMs in cyberspace do not focus on global common good creation but rather concern risk management. This is a remarkable limitation of scope, as the UN's remit is the good of all of humanity and the UN has a strong tradition to contribute to the global common good in other areas.

Still, weak links exist between UN work—the most recent consensus was limited to the UN GGE report of 2015—and global common good thinking. This is at the norms and principles level but not at the CBM level. For instance, the commitments expressed by the UN General Assembly to an open peaceful internet, international peace and security, to applying principles of humanity, and to consistency with the

UN Charter. The 2015 report also advised to explore in the future practical work, i.e. CBMs, such as developing common understandings on how international law applies for an open secure stable accessible and peaceful ICT environment, and conversely how what the concepts are of international peace and security in the use of ICTs area at technical, legal and policy level.

Clearly, the global common good approach to strategic autonomy, including for AI and cybersecurity deserves much more attention. It would be in the well-considered, self-interest of states for their sovereignty and in the interest of global business, it has a long tradition, and internationally the UN could give political support, and work with the private sector, internet community and civil society.

## 5 Policy Recommendations

Several of the policy recommendations that can be derived from the preceding analysis are not limited to AI. *Mutatis mutandis* the analysis is also applicable to other digital use cases and even to basic infrastructures such as electronic identification and authentication.

1. Risk management and resilience

   Governments, with the private sector, can promote the development and implementation of international norms and values to commit to 'do not harm' civilian infrastructure and for transparency. Practical work is needed in information exchange covering the whole AI chain, from pre-AI data capture to AI processing to post-AI explainability of algorithms. In addition, given the increasing sophistication of attackers, we need AI-enabled cyber exercises. We would also need interoperability, standardization, certification and promoting open data in sectoral cyber-resilience (to enable AI threat analytics fed by big data).

   Of high importance would be to rethink risk management rules and legislation in the perspective of human limitations relative to AI. With its micro-second speed deep analytics, AI is incomparably fast and powerful compared to the slow and difficult decision-making processes of traditional risk management. Yet it is these processes that are embedded into cyber resilience legislation such as the EU's Network and Information Security Directive. We need to know how to adapt legislation when de facto we have the 'human out of the loop' for AI-enabled critical infrastructure resilience.
2. Strategic partnerships

   For strategic partnerships it is important to reflect on the implications of AI and ethics guidelines (or law) applied to cybersecurity. The wisest route—and probably close to the heart of many—is to seek internationalizing such guidance as this would keep the global common good route open. The current push by the European Commission to pilot and validate its recent AI and ethics guidelines should therefore in the context of cybersecurity take the global dimension explicitly into account.

   There is also a need to address in defence and security policies the ethical certification for weaponized AI. At the same time, trade policy (such as the Was-

senaar agreement) must validate how export controls apply to cyber-related AI with ethical risks. In the European context this is part of the ongoing revision of export controls that brings in more explicitly the human rights dimension (European Commission 2016). An ambitious step, yet appropriate at UN level would be to discuss an international treaty for non-participation in AI cyber arms race and non-proliferation of cyber-arms (UNIDIR 2017 analyses challenges related to such an ambition).

3. Global common good

   Clearly this commentary makes an appeal to private–public collaboration and intergovernmental work at the UN to commit to pursuing global common goods for AI and ethics in cyberspace.

   In addition, states or alliances of states such as the EU can prioritize open source AI and AI for distributed security control in their R&D and investment policies.

   Finally, one of the most relevant and urgent global common goods to pursue is to ensure a proper interplay of AI and cybersecurity with personal data protection. As far as the EU is concerned in terms of AI and personal data the GDPR implementation should clarify the relationship between national security and transparency of automated decision-making.

## 6 Conclusions and Perspectives

We are addressing here the interplay of cybersecurity, AI and ethics in relation to sovereignty. Is there a conceptual framework that allows us combine multiple perspectives? I suggest that this would be a subject for further research. A few perspectives are provided in this concluding part.

We have analysed how sovereignty can be linked to three approaches to strategic autonomy and how these relate to ethics. Ethics as embedded in rules set by states relates to notions like ‘code’ conditions ‘law’ and vice versa. A link between cybersecurity and AI is that both are about intelligence: actors are intelligent, whether state and non-state real or virtual actors (embodied in AI).

Cybersecurity and ethics in relation to sovereignty are linked to the ethical consequence of the sovereignty gap, e.g. do we pass a threshold of harm such that addressing the sovereignty gap becomes a priority, i.e. when does state legitimacy get seriously affected?

This then leads us to investigate the ethical dimension of state legitimacy (Hurrell and Macdonald 2012). Internal legitimacy is the power and authority (implying a form of acceptance or consent) of the state and its institutions towards its citizens (Biersteker 2012). When the damage due to ‘cyber’ becomes too large authorities lose credibility. External legitimacy, the recognition of the final authority of a state by other states, is at risk when cyber vulnerabilities undermine the external credibility of a state’s power or when states are de facto subordinate to powerful global providers, like platform companies. Ethics here get close to sovereignty as a right, including a right to self-determination.

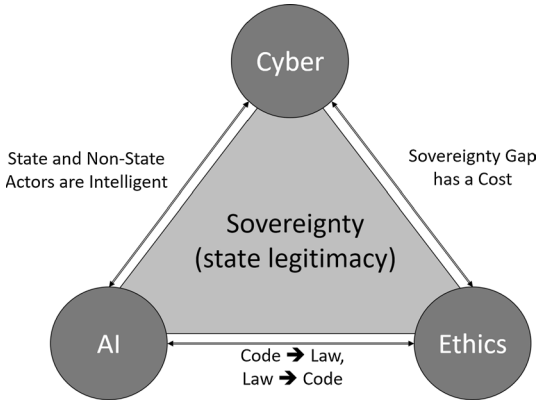


**Fig. 3** State legitimacy and ethics

This opens the debate about the primacy or cost of state sovereignty for example, relative to human rights which is an important element in the polarization on cybersecurity in the UN. Cost of legitimacy is then the notion underpinning the link between cyber and ethics in relation to sovereignty. Cyber raises that cost. The question is what the acceptable cost is of maintaining state sovereignty, i.e. what justifies plugging the sovereignty gap. This cost can include damage to people's life, such as not getting urgent healthcare (cf Wannacry) or suppression of freedom of expression (cf Uighur surveillance in China).

This gives us three conceptual links related to sovereignty (see Fig. 3): state and non-state actors are intelligent; the sovereignty gap has a cost; code conditions law and law conditions code. The focus of the debate then becomes (internal and external) state legitimacy which is a well-known notion in sovereignty political theory. State legitimacy is contestable by intelligent actors. Maintaining state legitimacy has a cost. State legitimacy is imposed on technology while technology also conditions state legitimacy. Given the challenges of AI and cybersecurity, a further reflection on ethics and state legitimacy may therefore be a fruitful area of research.



## References

Barlow, J.P. (1996). *A declaration of the independence of cyberspace*. Retrieved September 18, 2019, from https://www.eff.org/cyberspace-independence.

Biersteker, T. (2012). State, sovereignty and territory. In W. Carlsnaes, et al. (Eds.), *Handbook of international relations*. Thousand Oaks: SAGE Publications Ltd.

Broeders, D. (2017). Aligning the international protection of 'the public core of the internet' with state sovereignty and national security. *Journal of Cyber Policy, 2*(3), 366–376. https://doi.org/10.1080/23738871.2017.1403640.

Brundage, M., et al. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. Retrieved September 18, 2019, from https://maliciousaireport.com/.
Cowhey, P., & Aronson, J. (2017). *Digital DNA*. Oxford: Oxford University Press.
Drew, A., & Parton, C. (2019). *Committing to Huawei for 5G risks establishing a dependency*. Financial Times, Retrieved September 12, 2019.
European Commission. (2016). *Regulation setting up a Union regime for the control of exports, transfer, brokering, technical assistance and transit of dual-use items (recast).*
European Commission. (2019). *Ethics guidelines for trustworthy AI*. Retrieved September 18, 2019, from https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines.
Green, B. (2009). Lessons from the montreal protocol: Guidance for the next international climate change agreement. *Environmental Law*, *39*(1), 253–283.
Heinl, C. (2019). *CBMs How to build trust and confidence in cyberspace? Lessons and good practices—Cyber Direct Training*. European Institute of Security Studies/EU Cyber Direct, to be published.
Hurrell, A., & Macdonald, T. (2012). Ethics and norms in international relations. In W. Carlsnaes, et al. (Eds.), *Handbook of international relations*. Thousand Oaks: SAGE Publications Ltd.
Kello, L. (2017). *The virtual weapon and international order*. New Haven: Yale University Press.
Lessig, L. (2000). *Code is law*. Harvard Magazine 1 Jan 2000. Retrieved September 18, 2019, from https://www.harvardmagazine.com/2000/01/code-is-law-html.
Lessig, L. (2006). *Code: And other laws of cyberspace*. Basic Books (2nd ed.).
Maurer, T. (2018). *Cyber mercenaries: The state, hackers, and power*. Cambridge: Cambridge University Press.
Taddeo, M. (2018). Deterrence and norms to foster stability in cyberspace. *Philosophy & Technology, 31,* 323. https://doi.org/10.1007/s13347-018-0328-0.
Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science, 361*(6404), 751–752.
Timmers, P. (2019a). *Strategic autonomy and cybersecurity*. European Institute of Security Studies. Retrieved September 18, 2019, from https://eucyberdirect.eu/content_research/strategic-autonomy-and-cybersecurity/.
Timmers, P. (2019b). *Cybersecurity—Cyber direct training*. European Institute of Security Studies/EU Cyber Direct, to be published.
UNIDIR. (2017). The weaponization of increasingly autonomous technologies: Autonomous weapon systems and cyber operations. UNIDIR Resources, No. 7.