Chapter 8
Cybersecurity of Critical Infrastructure

Eleonora Viganò, Michele Loi, and Emad Yaghmaei

Abstract This chapter provides a political and philosophical analysis of the values at stake in ensuring cybersecurity for critical infrastructures. It presents a review of the boundaries of cybersecurity in national security, with a focus on the ethics of surveillance for protecting critical infrastructures and the use of AI. A bibliographic analysis of the literature is applied until 2016 to identify and discuss the cybersecurity value conflicts and ethical issues in national security. This is integrated with an analysis of the most recent literature on cyber-threats to national infrastructure and the role of AI. This chapter demonstrates that the increased connectedness of digital and non-digital infrastructure enhances the trade-offs between values identified in the literature of the past years, and supports this thesis with the analysis of four case studies.

Keywords Critical infrastructures · Cybersecurity · Ethical issues · National security · Value conflict

E. Viganò (✉)
Digital Society Initiative and Institute of Biomedical Ethics and History of Medicine, University of Zurich, Zurich, Switzerland
e-mail: eleonora.vigano@uzh.ch

M. Loi
Digital Society Initiative, University of Zurich, Zurich, Switzerland
e-mail: michele.loi@uzh.ch

E. Yaghmaei
Faculty of Technology, Policy and Management, Technical University of Delft, Delft, The Netherlands
e-mail: E.Yaghmaei@tudelft.nl; i.r.vandepoel@tudelft.nl

M. Christen et al. (eds.), *The Ethics of Cybersecurity*, The International Library of Ethics, Law and Technology 21,
https://doi.org/10.1007/978-3-030-29053-5_8

8.1 Introduction

One of the first duties of a national state is defending national security, which is the protection of its citizens, economy and institutions. Originally, national security pertained protection from military threats, but nowadays its scope is broader and includes security from terrorism and crime, security of economy, energy, environment, food, critical infrastructure, and finally cybersecurity. In this chapter, we tackle the ethical challenges posed by cybersecurity in national security and, in particular, the security of critical infrastructures. The critical infrastructures of a state are the physical, non-physical and cyber resources or services that are fundamental to the minimum functioning of a society and its economy. Reliable ICT networks and their services, which are critical infrastructures, are crucial in ensuring public welfare, economic stability, law enforcement and defence operations. Societies increasingly depend on public ICT networks and their services. The stability, safety and resiliency of the cyberspace is a national security issue, as the vulnerabilities of the cyberspace can be exploited to impair or destroy the critical infrastructures of a state, which highly rely on ICT networks and services.

In the national security sphere, state actors such as the police and national security agencies have privileged access to ICT services, in order to enforce the law and carry out defence operations and countermeasures to terrorism. However, the privileged access of government agencies to ICT services may endanger values that are pivotal for contemporary societies. Cybersecurity measures at the national level may create a condition of discrimination by affecting people's access to some resources or services, have economic implications that affect fairness, influence freedom of expression, limit people's autonomy and violate privacy (see also Chaps. 3 and 4). For this reason, the identification and discussion of the ethical issues and value conflicts involved in cybersecurity at the national level is fundamental to assist national security organisations. In this contribution, we answer this need by providing the main ethical issues and potential value conflicts that should be considered by every national security organisation when carrying out cybersecurity initiatives, with a specific focus on the vulnerabilities to which critical infrastructures are subject. The aim of this chapter is to raise awareness about cybersecurity values, and to stimulate idea generation and discussion regarding values of cybersecurity in the national security domain.

8.2 Review of the Literature on Cybersecurity in the National Security Domain

We identified the ethical issues at stake in cybersecurity in the national security domain in the papers selected in the literature review on cybersecurity and ethics by Yaghmaei et al. (2017). We then constructed a network of the ethical values involved and of their possible tensions within the network. As a starting point, we categorised

the papers by identifying value conflicts of cybersecurity initiatives. We further marked ethical issues and values that were either supportive or in conflict with security, as the latter is the core value of cybersecurity. On the basis of that categorisation, we delineated a set of ethical issues and conflicting values.

In our review of the papers on cybersecurity in the national security domain, two topics are mostly investigated. The first is the urgency for nations to develop strategies, frameworks, and suitable legal policies to defend and protect from cyber-attacks. The second topic is the difficulty and complexity of handling cyber-attacks countermeasures, which is because cyber-attacks overcome national borders and because interconnectivity, even though it boosts economic growth and makes people's life easier, nonetheless makes ICT networks and systems more vulnerable to attacks.

In the papers reviewed, cybersecurity is considered the top priority in dealing with terrorism and a necessary complement to national security strategies. Much of the literature indicates that national cybersecurity strategies need to be mindful of national cultures and ethical and technical values and at the same time compatible with international strategies and the global nature of the Internet.

The main ethical issues and conflicting values in national cybersecurity strategies that the authors of the reviewed papers have identified are shown in Table 8.1.

Table 8.1 The main ethical issues and value conflicts in the literature on national cybersecurity strategies

Ethical issue	Core value	Conflicting value
"Technology that was considered as a key contributor in progress of any country has evolved into a nightmare in form of cyber crimes" (Adeel et al. 2005)	Security (against cyber crime)	Connectivity
"Growing pressure for government to develop capacities to fight cyber wars" (Deibert 2011)	Security (against cyber terrorism/ cyber wars)	Protection of data
"Cyberspace enables cooperation and conflict in nearly equal measure" (Demchak 2011)	Security	Equity
"Focus on state's security crowds out consideration for security of an individual resulting in detrimental effect of the whole system" (Dunn Cavelty 2014)	Individual security	State security
"lawyers face dilemma because of the insufficient and vague cyber legislations are incompatible to deal with cyber crimes" (Faqir 2013)	Security (against cyber-crime)	Legality
"Infrastructure is owned and operated by private rather than public entities" (Hiller and Russell 2013)	Security	Surveillance
"Growth of criminal activities with the increased use of Internet and information technology" (Hui et al. 2007)	Security (against digital crime)	Accessibility
"Value of information increase so as well the efforts of criminals is more convenient" (Lehto 2013)	Security (against criminals)	Accessibility
"Information and communication technologies go beyond national boundaries" (Phahlamohlaka 2008)	Security	Protection of data

In the next sections, we provide a detailed list of ethical issues and conflicting values regarding cybersecurity in national security that were found and discussed in Yaghmaei et al. (2017).

8.2.1 Ethical Issues That Emerged in the Literature

Cyber Terrorism/Cyber Warfare Sekgwathe and Talib (2011: 171) argue that "Cyber-crime is typically understood to consist of accessing a computer without the owner's permission, exceeding the scope of one's approval to access a computer system, modifying or destroying computer data or using computer time and resources without proper authorisation. Cyber-terrorism consists essentially of undertaking these same activities to advance one's political or ideological ends." There is a twofold link between terrorism and the Internet. First, the Internet has become a forum for terrorist groups and individual terrorists, both to spread their messages of hate and violence, as well as to communicate with one another and their sympathisers. Second, individuals and groups have tried to attack computer networks, including those on the Internet; these acts are described as cyber terrorism or cyber warfare (Bucci 2012). Phahlamohlaka (2008) argues that the security risks associated with information and communication technologies, which go beyond national boundaries, are not fully in line with the value of data protection of all states. To avoid cyber warfare, the author contends that there is a need to develop and implement agile security-related ICT policies that mitigate the value conflict between data protection and security in the national security domain. Building on this value conflict, Deibert (2011) discusses the growing pressure on governments to develop capacities to fight cyber wars. He observes (2011: 1) that "today's deteriorating cyber-environment poses immediate threats to the maintenance of online freedom and longer-term threats to the integrity of global communications networks".

Cyber-Espionage Cyber espionage is the use of electronic capabilities to illegally gather information from a target. For all nations, the information technology revolution quietly changed the way governments operate. The asymmetrical threat posed by cyber-attacks and the inherent vulnerabilities of cyberspace constitute a serious security risk confronting all nations. The achievements of cyber espionage—to which law enforcement and counterintelligence have found little answer—hint that more serious cyber-attacks on critical infrastructures are only a matter of time (Geers 2010a). Nevertheless, national security planners should address all threats with method and objectivity. As dependence on IT and the Internet grows, governments should make proportional investments in network security and incident response to the cyber espionages (Geers 2010b; Lehto 2013).

Lack of Cyber Law The literature review reveals that legality problems play an important role in cybersecurity in the national security domain. Lawyers are faced

with insufficient and vague cybersecurity legislations, which are incompatible with the requirements for effectively dealing with cyber-crimes (Faqir 2013; see also Chap. 5), as we will see in the case study of Exodus in the final section of this chapter. At the same time, cyber laws have become more critical than before in data and information security, as one can see in the growth of cyber-criminal activities. Hui et al. (2007: 11) argue that "... digital crimes (e-crimes) impose new challenges on prevention, detection, investigation, and prosecution of the corresponding offences". Widely accessible systems should be made in a manner that enables one to detect and investigate digital crimes in a more efficient and effective way.

Cyber Awareness Raising awareness about cyber-security threats and vulnerabilities and their impact on society has become vital, but it seems to be missing in the society, if compared to the leadership that the governments of nations try to establish. By raising awareness, individual and corporate users can learn how to behave in the online world and protect themselves from typical risks. Awareness activities occur on an ongoing basis and use a variety of delivery methods to reach broad audiences. The awareness raising, however, varies across countries. Security awareness activities may be triggered by different events or factors, which may be internal or external to an organisation. Major external factors include recent security breaches, threats and incidents, new risks, updates of security policy and/or strategy. Examples of the internal factors are new laws and new governments.

Profiling In profiling, people are approached, judged or treated in a certain way because they have characteristics that fit a certain profile and are associated with certain other traits. Profiling is not addressed explicitly in the identified literature, but it is implicitly mentioned in four papers. Profiling is used for a wide range of purposes and by various actors. It is employed by police or security agencies to find criminals or terrorists, by airport security to decide whom to check more carefully, by companies to target certain consumers, and by banks in deciding to whom to give a loan. As these examples already suggest, sometimes profiling serves security objectives. At the same time, profiling may inflict all kinds of undeserved harm on people, from nuisance to false accusations to even, in extreme cases, unjustified imprisonment. Thus, profiling can create tension between values such as non-discrimination and absence of bias, on the one hand, and security, on the other. Although profiling may involve privacy violations—as personal information is gathered to fit somebody into a profile—the main issue at stake is not privacy. Rather, the issue is that a generalisation is made on the basis of limited information about a person. This generalisation is based on statistical information regarding a group to which a person belongs. However, in virtue of the probabilistic nature of such information, the latter may say nothing about a person. As a consequence, profiling may lead to stereotyping and discrimination, as has occurred in the use of facial recognition technologies by the police and security: such systems are less accurate for certain groups (Klare et al. 2012) and may lead to the discriminatory treatment of people (Introna and Wood 2004; Garvie et al. 2016), as we will see in the third case study that we present.

8.2.2 Value Conflicts Identified in the Literature

Privacy/Protection of Data ↔ Security A critical issue in cyberspace lies in the inability of companies and private businesses to exchange information with the government. This causes insufficient information collection, skews analysts' results, and prevents the states from collecting sufficient data on cyber-attacks and developing better defenses (McNally 2013). The cyber-attacks on Google illustrate the vulnerability of information stored in the cloud, online surveillance and private sector collaboration with government agencies against global terrorism. Hiller and Russell (2013) argue that cyber infrastructure is mainly owned and operated by private entities instead of public ones. Therefore, the states should select the most effective cybersecurity strategy and regulate the private sector to reduce overall cybersecurity risk and address the privacy concerns on cyberspace. We delve into this value conflict in the case study of Exodus. Furthermore, counter-terrorism measures and tools that tackle cyber-crime often invade privacy in the most brutal ways. At the same time, lack of personal online security leads to breaches of privacy. Security is thus an essential part of enabling privacy in the national security domain, especially with regards to data security, data protection, data ownership, access control, and information and computer security.

State Security ↔ Individual Security Dunn Cavelty (2014) discusses a lack of focus on individuals in the efforts of states to achieve security in the building of ICT and other critical infrastructures. As a result, he argues, state security is not aligned with individual security. In fact, the focus on state's security crowds out consideration for the security of individuals. The result is a detrimental effect of the whole system: the state actors militarise cyber-security and override the different security needs of individuals in the cyberspace.

Connectivity ↔ Security The urgency for nations to develop strategies, frameworks or suitable legal policies to defend and protect from cyber-attacks is discussed in several papers. At the same time, as mentioned, it is often contended that cyber-attacks beyond borders are increasingly difficult and complex to handle.

Accessibility ↔ Security With lower costs associated with information accessibility and retrieval, more consumers and producers have access to global markets and transnational communication. Many Internet users, however, are not fully aware of cyber threats and they are not trained to protect themselves against these threats, thus becoming vulnerable to online exploits and increasing insecurity in cyberspace.

Connectivity ↔ Equity of Access Globally interconnected digital information and communication underpin almost every facet of modern society and its critical infrastructure. However, not everyone in society has the same degree of access to information and communication technology. From the literature review, it emerged that

inclusion and equity of access, consumer and producer accessibility to global markets, transnational communication, learning, and entertainment should be guaranteed to all, without causing exclusion, along with connectivity.

Confidentiality ↔ Trust Confidentiality prevents the disclosure of information to unauthorised individuals or systems. The impact of cyber-threats could reduce public confidence and damage reputation of Internet transactions. Thus, assuring a trusted and resilient information and communications infrastructure is needed to protect privacy.

8.2.3 The Gap in the Literature

We observed that the examined literature fails to emphasise to a sufficient degree that cybersecurity in national security involves numerous conflicting values. By contrast, the literature generally tends to focus on only one value (e.g. security, privacy, connectivity). Moreover, two topics that are highly relevant for ethics in cybersecurity at the national level are overlooked in the articles we reviewed: limitation of democratic values and creation of power imbalances.

With regards to the risk that cybersecurity may limit democratic values, on several occasions, governments and security agencies have required access to encrypted communication such as that on WhatsApp for security reasons, e.g. to detect and avoid potential terrorist attacks. Opponents of such access do not only point to privacy considerations but also to the fact that encrypted communication that cannot be accessed by governments and their agencies might be important for the democratic process and support opposition movements in countries with totalitarian or suppressive regimes. A similar issue has arisen in relation to the Tor network. The latter is a free software and an open network that supports users in protecting themselves against traffic analysis, which is a form of network surveillance that threatens freedom and privacy. In the aftermath of the hacking of the Democratic Party during the U.S. elections, it transpired that a Dutch private Tor server had probably been used in the hacking. The Tor server was owned by Rejo Zenger, an employee of Bits of Freedom. Bits of Freedom is a Dutch digital rights organisation which focuses on privacy and freedom of communications in the digital age. Although Zenger recognises that Tor servers can be misused by hackers, and are in that sense a threat to cybersecurity, he believes that this is a price worth paying, not only for reasons of privacy but also because these servers may be crucial for whistle blowers to reveal abuses. Again, the value that is at stake here is not just privacy but also a range of civil liberties that are seen as crucial for democracy and the democratic process.

The second value issue that is neglected in the literature but relevant for cybersecurity in the national security domain regards economic and political power imbalances. Economic monopolies or oligarchies are often considered undesirable, and in democracies, the balance of the political power between citizens and their government is a fundamental goal. It is acknowledged that maintaining certain power

balances is important for a healthy economy and for democratic politics. What seems to be less recognised is that the possession of information about others and their behaviour is an increasing source of power in the information age. In fact, organisations that collect or possess large amounts of (personal) data may increasingly have power over other actors, which may lead to the disruption of existing power balances and the creation of new ones. The alteration of power balances pertains to companies such as Google or Facebook that collect large amounts of data about users and consumers, but also to governments and security agencies that may collect large amounts of data about citizens, and to providers of cybersecurity technologies, as these activities may involve accessing highly sensitive data. It should be noted that the accumulation of large amounts of data in the hands of a few may lead to power imbalances and may be problematic even if such data are anonymised, or if people have given their informed consent for the collection, storage, and use of their data. Consequently, even when privacy concerns are properly addressed, the accumulation of large amounts of data in the hands of a few may be considered problematic for economic as well as political reasons.

8.3 Cybersecurity of Critical Infrastructure

There are many definitions of critical infrastructures, which mirror cultural trends and historically evolving political needs (Office of the [US] President 2003; Federal Register 1996; Maglaras et al. 2018; Moteff and Parfomac 2004; Commission of the European Communities 2006). The common features of all these definitions include the idea that infrastructures are *general purpose means* to different kinds of human activities, in particular economic activities, but also activities necessary to protect security and health. One could compare critical infrastructures to the skull and bones of a body, to its blood vessels, to its nervous system: in short, to its vital organs, which need to be in place and work well for every action of the human body to be performed efficiently and painlessly.

Although nowadays all the systems that are comprised in critical infrastructure rely on ICT networks and services, they are not equally sensitive to attacks through cyber means. For example, hospitals and telecommunication systems, energy, banking and finance, and postal sectors, all rely on cyberinfrastructure to a such a degree that makes them obvious targets to an attacker.

We find that the definition of what counts as a cyber-attack to infrastructure is ambiguous, hence we introduce a classification of attacks by means of two orthogonal conceptual distinctions, leading to four distinct kinds of cyber-attacks to infrastructure. The types of attacks to critical infrastructure can be distinguished on the basis of the means of attack, as mere cyber-attacks vs. attacks with a physical component (physical or cyber-physical) and on the basis of the outcome damage, which can be physical (or physical and functional) vs. purely functional (see Table 8.2). We now describe the four possible combinations of means of attack and damage and all kinds of cyber-attacks.

Table 8.2 Types of attacks on critical infrastructure

Damage →	1. Physical or physical-functional	2. Merely functional
Means of attack ↓		
A. Physical or cyber-physical	A1	A2
B. Merely cyber	B1	B2

First, in terms of the damage caused by the attack, we can distinguish physical or physical-functional (1) from merely functional attacks (2). In our definition, when the attack is *merely* functional (2), the only object that gets destroyed is information. Although malfunctioning and disruption of services may follow from the attack, there is *no* physical damage. In a physical attack (1), the attacked object is "persons, property or infrastructure attacked *through* cyberspace" (Roscini 2017: 103). We can make this distinction more precise by appealing to a criterion that has been suggested in the law of armed conflict. According to this criterion, a cyber operation counts as a physical attack if "restoration of functionality requires replacement of physical components" (Schmitt 2013: 108). The criterion is controversial in its original legal function as a measure of attack severity legitimising a military response, because it treats as an attack the physical destruction of a single server but not the incapacitation of an object (e.g. civilian power station) for days (Roscini 2017). However, our question here does not concern the justification of acts of wars, thus the distinction is far less problematic in our context. We merely need it to rigorously distinguish purely functional (2) from physical attacks, which typically have *functional consequences* (thus the label physical or physical-functional, in 1). Any attack that causes physical damage to infrastructure belongs to the column 1, irrespective of the means of attack (which can be also be purely software-based, as in the Stuxnet case, see below).

Second, in terms of means of attack, we shall distinguish a 'merely cyber' attack (B), for example through a virus or trojan, from a physical attack (A). Ordinary physical attacks to physical infrastructure causing physical damage (A1), e.g. shooting a missile to bring down a bridge or throwing poison in the water pipes may not belong to the realm of *cyber*security. However, some such attacks do, for example, the use of drones hacked or guided by malicious AI to carry explosives in the proximity of a dam. An instance of A2 (physical attack without physical damage) can be the use of graphite bombs, which spread extremely fine carbon filaments over electrical components that cause fully recoverable physical damage to the infrastructure: a short-circuit and a disruption of the electrical supply (Roscini 2017). This clearly counts as a cybersecurity threat, and it may not count as a physical attack according to our definition, as it is possible that no physical component needs replacement. An example of B1 is Stuxnet, the virus targeting the Siemens software that operated the uranium enrichment facility in Iran, in which the attacked objects were the turbines themselves, not just the information in the system. In this case, the means of the attack, unlike the case involving drones, were merely informational (a piece of software), but the goal was to physically damage the turbines. Cell B2 comprises attacks that disrupt the informational infrastructure of a country, without

causing physical damage as defined. This includes, for example, DDoS attack that disrupt the processes of critical systems as well as the use of social media bots to spread dissent and convey political messages (Brundage et al. 2018). Any substantial and long perpetuated attack of the functioning of the Internet, when it does not cause physical damage to machineries or people, falls in category B2. An example is the sustained DDoS attack against the Chinese national domain name resolution registry on 25 August 2013, which interrupted or slowed down connectivity (Roscini 2017) without any lasting physical damage.

Therefore, the same critical infrastructure, e.g. the Internet, can be attacked by causing physical or merely functional damage, i.e. by targeting respectively its *hardware* or *software* components (Roscini 2017). The Internet is also vulnerable to both physical and 'merely cyber' means of attacks, e.g. missiles destroying servers and DDoS attacks, respectively. In *all* cases, the main impact on the population is that Internet connectivity is reduced, slowed down or made sloppy.

In all four kinds of attack to critical infrastructures, the vulnerable attack surface gets broader and broader due to digitisation—which means increased data availability and connectedness—and the development of AI—which obviously leads to augmenting the technological infrastructure for data collection and data analysis. We discuss two phenomena that are related to this issue, in the next section: first, the embedding of industrial control systems into public communication infrastructures. The traditional relative isolation and peculiar constitution of these information and communication systems has declined as business has turned to exploit peer-to-peer communications, real time monitoring, and lately, smart grids built through the Internet of things and other services provided through the Internet (Maglaras et al. 2018). This has implications for cybersecurity, as we will see. The second phenomenon is the diffusion of AI, which has three implications for the cybersecurity of national infrastructure. First, the widespread availability of new cyber-physical systems, which can be exploited by novel attacks, for example causing self-driving cars to crash (Brundage et al. 2018); this is typically a physical and functional attack; second, the vulnerability that follows from the embedding of AI in critical infrastructures itself, which makes them vulnerable to both functional and physical-functional (*à la* Stuxnet) attacks; third, the possibility of using AI to enhance the scale and/or sophistication of attacks (both purely cyber as well as cyber-physical) against the critical infrastructure itself.

8.3.1 Cybersecurity of Industrial Control Systems

The threat of cyber-attacks to infrastructure is capable of motivating the state to enhance its cyber capabilities. Unfortunately, some countermeasures of the state do not lead to enhancing the country's cyber defences directly, but rather enhancing investigative and retaliatory capabilities. State officials may recognise that there are structural limits that prevent improving the cyber defences of some critical

infrastructures to the degree needed by national security objectives, or at least, there are such limits for any society that is not ready to renounce the efficiency advances brought by increased connectedness through ICT and AI. As Maglaras et al. point out, these limits are due to the current industrial control system network, which is a "unique environment, that combines large scale, geographically distributed, legacy and proprietary system components" (Maglaras et al. 2018: 43). In a sense, the combination in the same network of ad hoc programmable logical controllers and proprietary systems (unconventional solutions) with well-documented protocols and off-the-shelf hardware solutions (conventional solutions) is the worst of all worlds from the point of view of cybersecurity. While unconventional solutions (which are still in place) may be poorly understood by cybersecurity specialists, the use of conventional ones threatens to undermine the obscurity of previous configurations, which are used to protect them from simple attacks (Maglaras et al. 2018). The combination of both solutions in the same network means that although the benefit of obscurity may be significantly reduced, it will still be very costly to guarantee high levels of security to such systems, as it requires ad hoc solutions.

The challenge in improving the strictly defensive cybersecurity programme of industrial control systems may lead, as a logical response by concerned politicians, to enhancing the capabilities of attack and surveillance by state agencies. This can be considered a strategy of *prevention* of attacks to critical infrastructure, and perhaps even *retaliation*, which appears all the more necessary since its *protection* is so challenging from a technical and financial perspective. The enhancement of *prevention,* which is achieved through surveillance, is, however, in a trade-off with citizen's privacy. The development of *retaliation* capabilities is in tension with the prospects of long-term cyber peace. Moreover, the technology risks escaping from direct control of the government and may create inequities in citizens' capacity to protect privacy and render privacy a luxury good. In other words, our hypothesis is that, considering national security as an integrated socio-technical system, the following socio-political chain (C) of events may be in place:

C1. Enhanced connectivity of critical infrastructure ➔ increased vulnerability of critical infrastructure ➔ increased political incentive to enhance prevention against internal (e.g. domestic terrorists) and external (e.g. enemy states) threats

Furthermore, the causal chain may continue in two distinct branches, one domestic and one that starts with foreign and may have domestic implications as well:

C1A. Increased political incentive to enhance prevention against internal threats ➔ greater threats to citizens' privacy and freedom ➔ increased inequity in the protection from surveillance

C1B. Increased political incentive to enhance prevention against external threats ➔ cyber-offensive capabilities to be used against foreign enemies ➔ increased distrust between states

C1B may in turn lead to a causal chain that reinforces the nefarious effects of C1A, namely:

C1C. Increased distrust between states ➔ development of cyber-offensive capabilities (e.g. zero-day exploits) ➔ possible misuse of cyber-offensive capabilities ➔ greater threats to citizen's privacy and freedom ➔ increased inequity in the protection from surveillance

In conclusion, there appears to be a trade-off between, on the one hand, the efficiency granted by embedding industrial control systems in larger and more general-purpose networks and by using off-the-shelf and more general-purpose information technology and, on the other, the capability to protect such systems from attacks. This conflict leads to further trade-offs if the states decide to protect infrastructure by developing preventive and retaliation offensive cyber capabilities.

8.3.2 AI and Cybersecurity of Critical Infrastructure

AI enhances the capabilities of attackers to affect the informational infrastructure of a society, as AI technologies are in general dual use (Brundage et al. 2018). For example, face-recognition and the ability to generate synthetic pictures and audios, or to manipulate existing ones, can be used to disrupt, among others, political processes. Recently, the literature on cybersecurity has turned its attention to the cyber vulnerabilities emerging from: (a) the increased use of AI in cyber-physical systems that, if hacked or repurposed, can pose novel threats to critical infrastructure; (b) the increased use of AI in critical infrastructure itself; and (c) the use of new AI-powered tools to launch more powerful attacks against critical infrastructure (Brundage et al. 2018).

An instance of (a) is the use of self-driving cars. Their AIs create opportunities for attacks through adversarial examples that cause crashes. If the attack is of sufficiently wide scope, it can be configured as an attack to a country's road networks, which are a critical infrastructure. Another example is the repurposing of commercial AI systems as physical weapons against infrastructure. For example, commercial drones and self-driving cars could be used to deliver explosives against physical infrastructures such as the electric grid, dams, hospitals, schools, etc. (Brundage et al. 2018). These attacks all fall into case A1 in our fourfold classification. Examples of type (b) derive from the fact that AI-augmented services are vulnerable to AI-specific attacks such as adversarial examples (Brundage et al. 2018). One case concerning a specific critical infrastructure, namely hospitals, is the possibility of adversarial attacks against diagnostic tools employing AI (Finlayson et al. 2019). These are instances of B2 in our classification. Finally, example of type (c) concerns the use of AI to enhance attacks against critical infrastructure. The autonomy of AI increases the potential damage that a single person may be able to cause (Brundage et al. 2018). The literature describes cases of both A1 and B2 cyber-attacks. Distributed attacks by networks of coordinated robotic systems (swarming attacks) such as drone swarms may be enabled by multi-agent

swarming networks, which are an instance of AI (Brundage et al. 2018). Face-recognition, navigation and planning algorithms are similar enhancements of robotic systems (Brundage et al. 2018), which can be used to launch physical attacks (A1) to infrastructures. Moreover, AI can be used to enhance the search of software vulnerabilities (Brundage et al. 2018; King et al. 2019), thus increasing the scale or sophistication of attacks to the software embedded in infrastructure. The effect can be functional disruption (B2) or physical damage (A1) when the infrastructure in question relies on information and communication technology for its functioning or safety.

In conclusion, the widespread availability of AI, which is a dual use technology, enhances the capabilities of attackers, by "alleviating the trade-off between scale and efficacy of attacks" (Brundage et al. 2018: 6) and by enabling new kinds of attacks, such as swarming attacks coordinated by AI frameworks.

8.3.3 Value Conflicts in the Use of AI in Cybersecurity in the National Security Domain

As discussed in the previous section, AI is taking both an attacking and defensive role in cybersecurity. One of the clearest demonstrations was the DARPA Cyber Grand Challenge of 2016, with AI systems able to both identify and patch vulnerabilities (King et al. 2019; Taddeo 2019). Some AI cybersecurity defences are familiar, such as spam filters and malware detectors. Other examples are defence drones and the use of AI in criminal investigations and terrorism (Brundage et al. 2018). The recent literature has identified three significant value conflicts concerning AI: (1) security vs. privacy, (2) non-discrimination vs. security, and (3) short-term security vs. long-term security in cybersecurity between nation states.

The first value conflict concerns the use of AI-empowered technology such as facial recognition or social network analysis (Brundage et al. 2018) for purposes of national security defence. The employment of AI in a defensive and preventive role may enable a faster identification and response to threats, but it will not protect society from the threat of authoritarian abuse of the cyber domain by states (Brundage et al. 2018). As AI is more pervasively used for image, video and text recognition by state agencies, the traditional trade-off of cybersecurity mentioned in Sect. 8.2.2 (*Privacy/Protection of Data ⇔ Security*) is exacerbated. Moreover, AI can be employed to better identify and profile citizens in relation to their online behaviour, for example through biometric profiles based on the way in which users move their mice (Taddeo 2019).

The conflict between non-discrimination and security is due to the biases and discriminations in AI, by which one means either *indirect discrimination/disparate impact*, which leads to certain groups (e.g. races, religions) being negatively affected by the outcome of the facially neutral algorithms, or *unequal accuracy*, which is the

different balance in false positive/false negative rates for different groups (Zafar et al. 2017; Chouldechova and Roth 2018). All kinds of systems employed for profiling dangerous individuals and predicting threats are affected by indirect discrimination and/or unequal accuracy. This is not due exclusively to biases in data collection, but also to unavoidable trade-offs between different kinds of biases (Chouldechova 2016; Kleinberg et al. 2016) and between bias-removal techniques and the accuracy or efficiency of the prediction, or classification, in question (Berk et al. 2017; Corbett-Davies et al. 2017). We examine a case study of the ethical conflict between non-discrimination and security in the next session.

The third value conflict is a tension between the short-term goal of enhanced security, which may be *also* promoted by cyber defences (Brundage et al. 2018), and the negative side-effects of such reliance in the long-term (Brundage et al. 2018; King et al. 2019; Taddeo 2019). While the current confidence of experts in these systems is low (Brundage et al. 2018), improving such systems has been recommended (Brundage et al. 2018), and it may be speculated that the AI testing of cybersecurity will greatly enhance cybersecurity and reduce the value of zero-day exploits (Taddeo 2019). Among the side-effects is, first, the fact that AI-based defences may also have unattended vulnerabilities (Brundage et al. 2018). Second, if AI testing of cybersecurity proves more accurate than the human testing in the short term, then a human deskilling problem follows, namely the risk that "delegating testing to AI could lead to a complete deskilling of experts [which] would be imprudent" (Taddeo 2019: 188). Third, there is the risk that AI-enabled cyber weapons will be used in national active cyber defence strategies, i.e. in order to retaliate or create deterrence (Taddeo 2019). Some scholars have argued that the use of AI-enabled cyber weapons by states, for purposes of retaliation and deterrence, will lead to a cyber arms race from which all involved parties will lose in terms of their national security (Taddeo and Floridi 2018). Thus, scholars have advocated the adoption of an international regime of norms regulating state behaviour in cyber space (Taddeo 2018; Taddeo and Floridi 2018). However, consensus on such norms for the specific case of AI is unlikely to be reached soon, witnessing the failure of governmental actors to agree on more general principles of cyberspace behaviour (see Chap. 18). For at least two decades, governments and scholars alike have been advocating a regime of responsible behaviour in cyberspace (see Chap. 18) of which norms concerning AI can be considered an extension. Similar proposals include common norms of collaboration and information sharing between states (see Chap. 13), in order to build and strengthen trust, and/or higher investments in the security and resilience of digital infrastructure, which reduce the benefit that can be derived from such attacks. In a similar vein, Lucas (in this volume) has placed emphasis on creating the conditions for the emergence of practices and customs that confer more stability and predictability of the behaviour of states in the cyber domain. This could be facilitated, he suggests, by promoting public-private partnership in cyberspace and investing in international cooperation, to identify malevolent cyber actors.

8.4 Case Studies of Cybersecurity in the National Security Domain

In what follows, we illustrate four case studies that are related to one or more ethical issues in cybersecurity at the national level that we tackled in this chapter. First, we present a case of cyber retaliation against a critical infrastructure, which threatens cyber peace (see also Chap. 13). Subsequently, we describe two cases of surveillance technologies that governments are pursuing to enhance their cyber capabilities, which may be misused against the governed. Finally, we address the case of some morally problematic cybersecurity threats exploited by governments against enemy states or internal opponents.

8.4.1 Iranian Attack to the US Power Grid System (Counter-Measure to Stuxnet)

In 2013, some hackers breached the control system of a dam near New York through a cellular modem and infiltrated the U.S. power grid system, gaining enough remote access to control the operations networks of the power system. The hackers targeted Calpine Corporation, a power producer with 82 plants operating in 18 states and Canada. Opening a pathway into the networks running the U.S. power grid was not difficult as the infrastructure was outdated and its ICT network was not sufficiently protected (Thompson 2016). Previously, various cyber-attacks from Russia and China to networks tied to the U.S. power grid were discovered, but in the case of the dam near New York, the hackers gathered much more data: passwords to connect remotely to the power grid's networks and detailed engineering drawings of networks and power stations from New York to California. Potentially they would have been able to shut down generating stations and cause blackouts, but their infiltration was discovered before they started damaging the power grid. The digital clues that were gathered pointed to Iranian hackers (Thompson 2016). In the same period, hackers linked to the Iranian government attacked American bank websites. These attacks were Iran's retaliation for Stuxnet.

It is likely that the infiltration into Calpine's network was part of the Iranian counter-attack and thus it can be considered a case of cyber warfare. The Calpine case shows that the exploit of vulnerabilities in the ICT systems by governments produces a cyber arms race. In fact, while the Stuxnet attack did not harm innocent civilians, the data gathered by the hackers attacking Calpine would have harmed civilians, if the plan had been completed. Furthermore, the aim of the Stuxnet attack was considered a worthy one by the majority of the international community, as it consisted in preventing Iran from acquiring nuclear weapons, even though it raised several moral concerns (Baylon 2017). A final ethical issue that characterises the Calpine case is the tension between resource investment and security: enhancing the network security of energy infrastructures is a costly operation that requires significant investments.

8.4.2 Hacking of Citizens' Telephone with Exodus

In many countries in Europe and in the U.S., law enforcement and investigation can legally hack the devices of targets if required by a court order. In Italy, the police used Exodus, which is a spyware for smartphones, to gather data from criminals' cell phones (e.g. their telephone book, call and browsing history, GPS position, text messages, audio recordings of the phone's surroundings, etc.) and to send commands to the infected cell phone via a port and a shell. Exodus was uploaded in more than 20 Android applications on the official Google Play Store, which were mostly apps to receive promotions and marketing offers or to improve the smartphone's performance. Thus, these apps attracted and were downloaded by innocent people. Their phone was infected because Exodus installed itself on any phone without validating that the target was legitimate, whereas it should have checked the devices' IMEI to verify if the phone was intended to be targeted. Moreover, the port that was opened by Exodus could be exploited by anyone on the same Wi-Fi network, thus enabling the hacking of the infected phone to third parties. Google declared that less than 1000 mobile phones of Italian customers were infected (Franceschi-Bicchierai and Coluccini 2019).

In such a case we see, first, the opposition between national security in the form of the fight against crime, which is the aim pursued by the Italian state police and magistrates, versus the practical realisation of this aim. The latter involved innocent people and the violation of their privacy for no legitimate reason, since they were not under investigation. Furthermore, these people were rendered more vulnerable, as following the infection their mobile phone could be hacked by potentially everyone. Second, we observe a tension between legality and security, as the Italian legal framework on cybersecurity is not keeping pace with the new technologies adopted in criminal surveillance. The 2017 Italian law regulating legal spyware and its 2018 integration are too vague and do not address the need to protect the overall security of a targeted telephone. The results of such legal framework is that Exodus could be equated with old physical surveillance devices such as hidden microphones, whereas it is much more invasive (Franceschi-Bicchierai and Coluccini 2019). The society that the State police hired to develop Exodus is to be held responsible for infecting non-targeted people, as it deliberatively uploaded the apps with Exodus on Play Store, most likely in order to use innocent customers as oblivious experimental subjects for its software. Thus, it is likely that Exodus's failure to check the target's IMEI was not a programming error. Finally, Apple adopts filters that prevent malware from slipping onto its store that are stricter than those employed by Google. Apple's higher level of control protects its customers but has repercussions on the prices of Apple devices. This means that citizens' privacy is not equally protected: citizens with more economic resources can afford Apple's devices and be more protected.

8.4.3 'Biased' Face Recognition Systems

Face recognition systems (FRSs) are software used by police departments and airport security to respectively identify suspects and collect information regarding passengers with criminal records. The main reason why FRSs are increasingly employed by State agencies is that the task of finding a 'face in the crowd' or identifying a suspect from pictures of known offenders is a difficult task that requires effort. The FRSs automate this task and thus free government employees for more valuable tasks. FRSs are highly desirable as a biometric for digital surveillance as they are silent, non-invasive, and above all they are the only biometric techniques currently used by law enforcement that do not require the explicit consent of the subject. However, the performance of FRSs is highly reduced in an uncontrolled 'face-in-the-crowd' environment, in the case of a large database, and if there is an elapsed time between the database image and the probe image (Introna and Wood 2004).

The first ethical issue raised by the implementation of FRSs in general is the reduction of citizens' privacy, as FRSs can use the data from any CCTV camera system, for the sake of security. The second ethical issue is that FRSs were found to have lower performances on certain demographic groups: females, Afro-Americans, and young people (Klare et al. 2012), thus generating a form of discrimination. In the U.S., the criminal justice system and law enforcement are already affected by racial disparities, as black people are more scrutinised than white people by the police. FRSs may exacerbate this disparity as they increase the frequency that an innocent Afro-American suspect will come under police scrutiny (McCullon 2017). FRSs are increasingly employed by state agencies even because they should not be subject to the biases of human vision; they should be neutral, as they are technological artefacts. However, they are designed by humans in a specific sociotechnical context. This means that the biases of the algorithms of FRSs can be present in every phase of the algorithm design, from the selection of the data to the translation of the goal of the algorithm into mathematical constructs, to the selection of the tests that verify the performance of the algorithm (Loi et al. 2019). Hence, intentional attention to fairness in algorithm design is required for systems to overcome human biases and really achieve the equal treatment of individuals before the law.

8.4.4 Government Buying Zero-Day Exploits

Nowadays, cyber warfare comprises the practice of government agencies in buying zero-day exploits in the grey market. Prominent buyers of zero-day exploits are the governments of the U.S., Brazil, U.K., India and Israel. As these transactions occur in the grey markets and governments buy them in order to attack other countries or

opponents, these purchases are secret, and mentioning a specific real case is not possible. However, it is possible to delineate the dynamics of such transactions, thanks to the disclosures of hackers trading with government agencies (Perlroth and Sanger 2013).

The zero-day exploits can be used as a form of weapon, as they can disrupt and destroy computers and their network. The targets can be critical infrastructure and services vital to the economy, public health and national security of a country. Government buying vulnerabilities protect their national security by threatening that of other countries. The paradoxical consequence is that if each government seeks the vulnerabilities of the other governments in order to protect itself, in the long run each one will be less secure. This practice is an instance of the conflict between short-term security and long-term security (the third value trade-off of AI in national cybersecurity). The zero-day exploits can also be used by governments to monitor the activity of political dissenters, thus violating the privacy of these persons. The zero-day exploits *per se* are not harmful (Dunn Cavelty 2014); it is the purpose of their use that can be moral or immoral. A further ethical tension regarding governments buying vulnerabilities is between the hackers' business aim to maximise profits and the government's duty to ensure adequate cyber defence (Baylon 2017). Furthermore, cybersecurity should be a public good, but the governments buying zero-day exploits have to follow the logic of market. Lastly, as zero-day exploits are kept secret, they may benefit few people and empower institutions that are already powerful.

8.5 Conclusion

This chapter provided a political and philosophical analysis of the values at stake in ensuring cybersecurity for critical infrastructure. We applied a bibliographic analysis of the literature until 2016 to identify and classify cybersecurity value conflicts and ethical issues in national security. We then interpreted the recent literature as suggesting that the increased connectedness of digital and non-digital infrastructure enhances the trade-offs between the values we identified in the literature of the past few years. This is due primarily to two phenomena: first, the embeddedness of an individual control system in conventional networks and technological solutions and, second, the diffusion of AI, which broadens the attack surface (e.g. self-driving cars and other robots) and enhances the capabilities of hackers and crackers. We presented four case studies that show the trade-offs involving security in cybersecurity at the national level—which is the core value of cybersecurity—and the values that most frequently conflict with that: non-discrimination, equity, privacy, and long-term security.

Acknowledgments The chapter was created with funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 700540 and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 16.0052-1.

References

Adeel M, Chaudhry A, Shaikh R et al (2005) Taxonomy of cyber crimes and legislation in Pakistan. In: Proceedings of 1st international conference on information and communication technology, ICICT 2005, p 350

Baylon C (2017) Lessons from Stuxnet and the realm of cyber and nuclear security: implications for ethics in cyber warfare. In: Taddeo M, Glorioso L (eds) Ethics and policies for cyber operations. Springer, Cham, pp 213–229. https://doi.org/10.1007/978-3-319-45300-2_12

Berk R, Heidari H, Jabbari S et al (2017) A convex framework for fair regression. ArXiv:1706.02409. http://arxiv.org/abs/1706.02409. Last access 7 July 2019

Brundage M, Avin S, Clark J et al (2018) The malicious use of artificial intelligence: forecasting, prevention, and mitigation. ArXiv:1802.07228. http://arxiv.org/abs/1802.07228. Last access 7 July 2019

Bucci S (2012) Joining cybercrime and cyberterrorism: a likely scenario. In: Reveron DS (ed) Cyberspace and national security: threats, opportunities, and power in a virtual world. George Town University Press, Washington, DC, pp 57–68

Chouldechova A (2016) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. ArXiv:1610.07524. http://arxiv.org/abs/1610.07524. Last access 7 July 2019

Chouldechova A, Roth A (2018) The frontiers of fairness in machine learning. ArXiv:1810.08810. http://arxiv.org/abs/1810.08810. Last access 7 July 2019

Commission of the European Communities (2006) Communication from the Commission on a European Programme for Critical Infrastructure Protection. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2006:0786:FIN:EN:PDF. Last access 7 July 2019

Corbett-Davies S, Pierson E, Feller A et al (2017) Algorithmic decision making and the cost of fairness. ArXiv 1701.08230. https://doi.org/10.1145/3097983.309809

Deibert R (2011) Tracking the emerging arms race in cyberspace. Bull At Sci 67(1):1–8. https://journals.sagepub.com/doi/pdf/10.1177/0096340210393703

Demchak CC (2011) Wars of disruption and resilience: cybered conflict, power, and national security. University of Georgia Press, Athens

Dunn Cavelty M (2014) Breaking the cyber-security dilemma: aligning security needs and removing vulnerabilities. Sci Eng Ethics 20(3):701–715

Faqir RSA (2013) Cyber crimes in Jordan: a legal assessment on the effectiveness of information system crimes law no (30) of 2010. Int J Cyber Crim 7(1):81–90

Federal Register (1996) Executive order 13010 – critical infrastructure protection. 61(138): 37347–37350

Finlayson S, Bowers JD, Ito J et al (2019) Adversarial attacks on medical machine learning. Science 363(6433):1287–1289. https://doi.org/10.1126/science.aaw4399

Franceschi-Bicchierai, L, Coluccini R (2019, March 29) Researchers find Google Play Store Apps were actually government malware. Vice. https://www.vice.com/en_us/article/43z93g/hackers-hid-android-malware-in-google-play-store-exodus-esurv. Last access 7 July 2019

Garvie C, Bedoya AM, Frankle J (2016, October 18) The perpetual line-up. Unregulated police face recognition in America. Georgetown Law Center on Privacy & Technology. https://www.perpetuallineup.org. Last access 7 July 2019

Geers K (2010a) The challenge of cyber attack deterrence. Comput Law Secur Rev 26(3):298–303

Geers K (2010b) The cyber threat to national critical infrastructures: beyond theory. J Digit Forensic Pract 3(2/4):124–130

Hiller JS, Russell RS (2013) The challenge and imperative of private sector cybersecurity: an international comparison. Comp Law Secur Rev 29(3):236–245

Hui LCK, Chow KP, Yiu SM (2007) Tools and technology for computer forensics: research and development in Hong Kong. In: Dawson E, Wong DS (eds) Information security practice and experience, ISPEC 4464, pp 11–19

Introna L, Wood D (2004) Picturing algorithmic surveillance: the politics of facial recognition systems. Surveill Soc 2(2/3):177–198

King TC, Aggarwal N, Taddeo M et al (2019) Artificial intelligence crime: an interdisciplinary analysis of foreseeable threats and solutions. Sci Eng Ethics:1–32. https://doi.org/10.1007/s11948-018-00081-0

Klare BF, Burge MJ, Klontz JC et al (2012) Face recognition performance: role of demographic information. IEEE Trans Inf Forensics Secur 7(6):1789–1801

Kleinberg, J, Mullainathan S, Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. ArXiv:1609.05807. http://arxiv.org/abs/1609.05807

Lehto M (2013) The ways, means and ends in cyber security strategies. In: Kuusisto R, Kurkinen E (eds) Proceedings of the 12th European conference on information warfare and security, pp 182–190

Loi M, Ferrario A, Viganò E (2019) Transparency as design publicity: explaining and justifying inscrutable algorithms. SSRN scholarly paper ID 3404040. https://doi.org/10.2139/ssrn.3404040

Maglaras LA, Kim K, Janicke H et al (2018) Cyber security of critical infrastructures. ICT Express 4(1):42–45. https://doi.org/10.1016/j.icte.2018.02.001

McCullon R (2017, May 17) Facial recognition technology is both biased and understudied. Undark. https://undark.org/article/facial-recognition-technology-biased-understudied/. Last access 7 July 2019

McNally J (2013) Improving public-private sector cooperation on cyber event reporting. In: Hart D (ed) Proceedings of the 8th international conference on information warfare and security, pp 147–153

Moteff J, Parfomac P (2004) Critical infrastructure and key assets: definition and identification. Congressional report ADA454016. Library of Congress Washington DC Congressional Research Service. https://apps.dtic.mil/dtic/tr/fulltext/u2/a454016.pdf. Last access 7 July 2019

Office of the [US] President (2003) The National strategy for the physical protection of critical infrastructure and key assets, US White House Office. https://www.hsdl.org/?view&did=1041. Last access 7 July 2019

Perlroth N, Sanger DE (2013, July 13) Nations buying as hackers sell flaws in computer code. The New York Times. https://www.nytimes.com/2013/07/14/world/europe/nations-buying-as-hackers-sell-computer-flaws.html. Last access 7 July 2019

Phahlamohlaka J (2008) Globalisation and national security issues for the state: implications for national ICT policies. In: Avgerou C, Smith ML, van den Besselaar P (eds) IFIP international conference on human choice and computers, Social dimensions of information and communication technology policy 282, pp 95–107

Roscini M (2017) Military objectives in cyber warfare. In: Taddeo M, Glorioso L (eds) Ethics and policies for cyber operations: a NATO cooperative cyber defence centre of excellence initiative, Philosophical studies series. Springer, Cham, pp 99–114. https://doi.org/10.1007/978-3-319-45300-2_7

Schmitt MN (2013) Tallinn manual on the international law applicable to cyber warfare. Cambridge University Press, Cambridge

Sekgwathe V, Talib M (2011) Cyber crime detection and protection: third world still to cope-up. In: Yonazi JJ, Sedoyeka E, Ariwa E, El Qawasmeh E (eds) e-Technologies and networks for development, Communications in Computer and Information Science, ICeND 2011 171, pp 171–181. https://link.springer.com/chapter/10.1007/978-3-642-22729-5_15

Taddeo M (2018) Deterrence and norms to foster stability in cyberspace. Philos Technol 31(3):323–329. https://doi.org/10.1007/s13347-018-0328-0

Taddeo M (2019) Three ethical challenges of applications of artificial intelligence in cybersecurity. Mind Mach 29(2):187–191. https://doi.org/10.1007/s11023-019-09504-8

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

Thompson M (2016, March 26) Iranian Cyber Attack on New York Dam shows future of war. Time. https://time.com/4270728/iran-cyber-attack-dam-fbi/. Last access 7 July 2019

Yaghmaei E, Van de Poel I, Christen M (2017) Canvas white paper 1 – cybersecurity and ethics, SSRN scholarly paper ID 3091909. Social Science Research Network, Rochester. https://papers.ssrn.com/abstract=3091909

Zafar M, Bilal H, Valera I et al (2017) Fairness beyond disparate treatment & disparate impact: learning classification without disparate mistreatment. ArXiv 1610.08452:1171–1180. https://doi.org/10.1145/3038912.3052660