AI in Cyber Operations: Ethical and Legal Considerations for End-Users

Kirsi Helkala, James Cook, George Lucas, Frank Pasquale, Gregory Reichberg and Henrik Syse

Abstract This chapter addresses several of the principal ethical challenges confronting current and proposed new uses of artificial intelligence (AI) undertaken to enhance cybersecurity and conflict operations. Current uses of ”modular” AI - namely AI tailored for specific tasks and involving closely-supervised machine learning - already pervade both defensive and offensive cyber operations with varying degrees of efficiency and success. Representative samples of these are illustrated and discussed, highlighting some of the moral challenges accompanying each. The chapter then concludes with proposed uses of general AI, including ”deep learning” enabling expanded autonomy, along with the attendant risks and prospective moral dilemmas arising from such uses within the cyber domain. Our specific focus will be the impact, both favorable and unfavorable, on the character and capacities of the individuals and teams of professional experts who will constitute the end-users of AI-enhanced cybersecurity measures.

Kirsi Helkala
Norwegian Defence University College/Cyber Academy, Lillehammer, Norway;
Peace Research Institute Oslo, Oslo, Norway, e-mail: khelkala@mil.no

James Cook
United State Air Force Academy, Colorado Springs, CO, USA, e-mail: james.cook@usafa.edu

George Lucas
United State Naval Academy, Annapolis, MD, USA, e-mail: george.r.lucas.jr@gmail.com

Frank Pasquale
Brooklyn Law School, Brooklyn, NY, USA, e-mail: frank.pasquale@gmail.com

Gregory Reichberg
Peace Research Institute Oslo, Oslo, Norway, e-mail: greg.reichberg@prio.org

Henrik Syse
Peace Research Institute Oslo, Oslo, Norway, e-mail: syse@prio.org

1 Introduction

In December 2020, the European Union's Agency for Cyber Security released a detailed assessment of the security threats posed by artificial intelligence [2]. The report noted that AI and cybersecurity "have a multi-dimensional relationship and a series of interdependencies." The authors identified three broad threat categories: (1) providing cybersecurity for AI itself when used in a range of other operations (e.g., autonomous combat weapons and logistical systems, driverless cars, and a range of devices connected within the Internet of Things); (2) using artificial intelligence to enhance cybersecurity operations, both offensive and defensive; and (3) the malicious use of AI by adversaries to create and launch ever more sophisticated, malicious, and destructive cyberattacks.

This chapter focuses primarily upon the ethical and legal challenges posed by pursuing the *second* of these threat categories. Here we consider only cyber operations that are conducted and defended within cyberspace exclusively. This means that we exclude from this chapter the social and cognitive domain operations that also are enabled by cyber measures. Invoking this demarcation implies, for example, that discussion of malware incidents such as "SolarWinds" or "NotPetya" would fall within the purview of our study as having transpired exclusively within the cyber domain, while campaigns of disinformation, for example, using social media to disrupt real-world political activities, would not. We recognize the latter category of operations as equally significant, but perhaps better suited for treatment in a separate discussion. More specifically, as NATO alliance members alongside the USA and the nations of the European Union seek to enhance their own individual and collective security through such measures, the authors of this chapter examine the ethical and legal concerns they may well confront, and how these might be resolved to the satisfaction of the controlling governments and societies. In particular, we consider which tactical alternatives pose the least risk of legal liability or moral injury to the individual citizens of those nations and alliances who will be most closely involved in designing and carrying out these AI-enhanced cyberspace operations.

We distinguish between two distinct categories of ethical challenges: those arising from deliberate or wantonly reckless misuse of such technology, and a wholly separate and more widespread admonition that end-users attempt, in good faith, to guard against carelessness, negligence, or the emergence of unintended consequences arising from their otherwise justifiable and presumably beneficial uses of AI.

2 Background: AI and cyber operations

2.1 Artificial intelligence

There are several definitions of artificial intelligence, no one of which is definitive. High-level doctrine in the European Union [3] defines AI systems as

> software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal.

NATO has not issued joint doctrine pertaining specifically to use of AI in cyberspace. However, NATO's "Allied Joint Doctrine for Cyberspace Operations" [4], cited extensively below, in all relevant respects appears fully compatible with the EU doctrine ([3], p.1.), in which the high-level description of the operations and capacities of AI systems

> either use symbolic rules or learn a numeric model, and . . . can also adapt their behaviour by analysing how the environment is affected by their previous actions.

As a scientific discipline, AI includes several approaches and techniques [5], including various degrees of machine learning (ML) (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems).

Currently AI applications are all around us; we use them on a daily basis whether we are aware of them or not. Easy examples are online searches, product recommendations, and the vast array of algorithms enabling "self-driving" cars and similar machines. In cybersecurity, common examples are malware detection and classification, network intrusion detection, file type, network traffic and SPAM identification, insider threat detection and user authentication [6, 7].

These AI-supported applications are examples of modular AI [1] ("narrow" or "weak" AI are alternatives that frequently appear in the literature on AI). This mode of AI can be phenomenally good at single tasks in strictly defined environments - like winning at chess or Go; but it is useless where more general intelligence is needed in broader contexts and unfamiliar environments [8]. General artificial intelligence (often termed "strong AI") is closer to human intelligence in operating across a variety of cognitive tasks. A key element of general AI is the ability to improve its knowledge and performance on its own. Unsupervised ML systems, for instance, do not merely follow a pre-determined algorithm. Instead, they generate their own algorithms from information they gather. While general AI has not yet been attained in any of its envisioned forms (let alone embedded in cybersecurity systems at present), there are a variety of approaches that fall somewhere between modular and general AI, reflecting distinct paradigms of ML [9, 10].

AI systems (including those used to provide defensive cybersecurity) are vulnerable to attacks, failures, and accidents. The list of the possible threat actors conducting AI adversarial attacks is the same for traditional cyberattacks; these actors vary from the unsophisticated (script kiddies) to the sophisticated (state actors and state-paid actors), including cyber criminals, terrorists, hacktivists, as well as malicious and non-malicious insiders [2].

The European Union's Agency for Cyber Security categorizes threats into eight main categories [2]: *nefarious activity/abuse*, namely malicious action to disrupt the functioning of AI systems; *surveillance/interception/hijacking* involves actions whose goal is to eavesdrop or otherwise control communications; *physical attacks* sabotage the components or infrastructure of the system; *intentional damages* are accidental actions harming systems or persons; *failures/malfunctions* happen when a system itself does not work properly; *outages* are unexpected disruptions of service; *disasters* are natural catastrophes or larger accidents; and *legal threats* occur when a third party uses applicable law, domestic or international, to constrain the cyber operations of an opponent.

Adversarial input attacks and poisoning attacks [11, 12] are examples of abuse. In input attacks, the input is manipulated in ways that may go unnoticed by the human eye, as, for instance, when tape is affixed to a road sign [11]. Poisoning attacks, by contrast, target the design process of AI and the learning process itself. Poisoning can be done both to algorithms and to the massive data sets upon which machine learning is based [11]. In addition to nefarious activity, Hartmann and Steup [12] discuss techniques that can be used in eavesdropping on different neural networks and support-vector networks. The entry and exit points (not all applicable for each method) include input, labelling, pre-processing, feature extraction, classifiers, and weights. Noise generator and selector points are included simply to illustrate that AI methods are vulnerable both to actions with unintentionally destructive consequences and to deliberately malicious actions. When data sets are corrupted or algorithms are biased, the conclusions AI reaches will likewise be distorted or biased. And here lies the core of the problem: to what extent can we trust the results and decisions recommended by AI?

2.2 Cyberspace operations

AI solutions vary in complexity, and the context where they are used sets unique demands regarding both the requirements for flawlessness of decisions and the degree of autonomy delegated to AI. This applies especially with respect to AI tools used in cyberspace. NATO, the USA, and allies (such as members of the "Five Eyes") currently use AI to shrink critical timelines for cyber threat situational awareness. AI use in cyber operations enhances the capability to detect threats and malicious activities at a rate that is not humanly possible. Suspicious events, behaviors, and anomalies can be rapidly identified for cyber professionals and operators to further investigate and deploy mitigation strategies. This decreases the likelihood of adver-

saries gaining access to NATO and US-DoD networks, infrastructure, and weapon systems [13].

According to NATO's operational doctrine for cyberspace ([4], p.2),

> Cyberspace is not limited to, but at its core consists of, a computerized environment, artificially constructed and constantly under development.

NATO doctrine ([4], p.3) divides cyberspace into three layers: physical (e.g., hardware), logical (e.g., firmware, protocols) and cyberpersona (virtual identities, e.g., emails, net-profiles). The logical layer is always involved in cyberspace operations (COs), but effects of those operations can also impact both the physical and cyberpersona layers of cyberspace. Furthermore, NATO doctrine recognizes that COs can affect human senses, decision-making, and behavior. Similarly, COs can have an impact on other physical elements that are directly included within or connected to cyberspace. Activities outside of cyberspace that have an effect on cyberspace (e.g., the physical sabotage of hardware components), are not, however, considered COs.

In general, the basic principles applicable to NATO joint operations apply to cyberspace operations just as they do to operations in the kinetic domain (a separate document governs targeting in this domain). However, the concept of time and reach can be somewhat different in different situations. Cyberspace operations doctrine [4] lists some direct and indirect effects that cyberspace operations can have. The first five (1.-5.) are effects that defensive operations normally have in one's own communication and information systems (CIS), while the remaining (6.-10.) are effects that offensive operations can have on the other's (adversary's) network.

1. **Secure** against compromise of CIA (confidentiality, integrity and availability) of our own CIS, as well as the data they store.
2. **Isolate** the communication between adversaries and our own affected systems.
3. **Contain** the spread of the malicious activity.
4. **Neutralize** malicious activity permanently from our own CIS.
5. **Recover** quickly from the effects of malicious activity (network resilience).
6. **Manipulate** the integrity of an adversary's CIS.
7. **Exfiltrate** the information of adversaries through unauthorized access to their own CIS.
8. **Degrade** an asset of an adversary to a level below its normal capacity or performance.
9. **Disrupt** an asset of an adversary for an extended period of time.
10. **Destroy** an asset of the adversary.

From a legal perspective, cyber operations are expected to conform to international law (such as the United Nations Charter, Laws of Armed Conflict and human rights law) as well as to applicable domestic law (i.e., the laws of the nation carrying out COs). A specific cyber operation plan should include a description of the rules of engagement (ROE), as well as define the standing authority and expected effects of COs. Estimation of effects on dual-use objects (e.g., network infrastructure), other prospective collateral damage, as well as likelihood of attribution for the operation can be difficult [4] to determine.

Instruments of international law, as mentioned above, are extrapolated to cyberspace operations in the *Tallinn Manual 2.0* [14], while treaties and legislation pertaining to conventional armed conflict were extrapolated to the cyber domain in the initial *Tallinn Manual 1.0* [15]. The manuals are not presented as binding international law, but like many other manuals (e.g., the *San Remo Manual on International Law pertaining to Armed Conflicts at Sea* [16]), are intended to provide guidance in applying existing law to specific complex or novel situations. In this chapter, we purposely leave out of the discussion the specifics on who in the organization has responsibility for conducting such operations, and which type of cyber operations are intended, as these specifics obviously vary from one state jurisdiction to another.

2.2.1 Some examples of AI-supported cyberspace measures

A useful listing of defensive and offensive COs can be found in Truong, Diep and Zelinka [17]: On the defensive side of AI-based cyber applications, they list malware detection (PC malware, Android malware), network intrusion (intrusion detection, anomaly detection), phishing/spam detection (web phishing, mail phishing, spam on social media, spam mail) and other, similar measures (countering advanced persistent threats (APTs), identifying domain names generated by domain generation algorithms). For malicious use of AI, they list autonomous intelligent threats (strengthening malware and social engineering) and tools for attacking AI models (adversarial inputs, poisoning training data and model extraction).

Kaloudi and Li [18] have made a similar survey to the one above, but they focus chiefly on the offensive side with examples that they also map onto the cyber targeting chain that is often used when attack vector vulnerabilities are estimated. This chain contains three main phases and seven subphases: the *planning* (reconnaissance and weaponization), *intrusion* (delivery, exploitation, and installation) and *execution* phase (C2 and actions). AI methods are used in the *reconnaissance* phase for selecting targets and learning targets' standard behavior. In the *weaponization* phase, AI can be thought to generate attack payloads, aid password guessing or brute force attacks, generate abnormal behavior and find new vulnerabilities. In the *delivery* phase, AI methods help attacks to remain undetectable and conceal malicious intent. Automated methods establish means of distributing disruptive content in the *exploitation* phase. AI methods that can evolve and self-propagating malicious code are both used in the *installation* phase. In the *C2* phase and *action* phase, AI activates the malicious code and harvests the outcomes of the actions. Kaloudi and Li also discuss using AI-based methods on the defensive side. In their framework, methods for behavioral and risk analysis are to be used in the *planning* phase. Methods that are suited to detect anomalies and offensive AI patterns are suited for the *intrusion* phase. In the *execution* phase, AI methods should handle real-time response and configuration management.

Some AI methods used in real-time response against cyberattack fall within a "gray zone" between defensive and offensive tools. James Pattison [19] speaks of active cyber defense when the organization that is first to be attacked attacks or hacks

back. Pattison's contrast appears to discriminate between "offensive measures" as a tactic of "active defense" (i.e., both repelling an initial attack and simultaneously taking up the initiative of retaliation within the initial attack framework), and the more conventional meaning of "offensive" as "initiating the conflict" (strategic offense). The definition of passive and active cyber defense is based on where the action occurs. When the cyber measures are initiating activity only inside the targeted organization's network, they are called passive defensive measures (e.g., firewalls). If the defensive activity extends beyond the targeted organization's network, it is classified under active defensive measures. The disruptiveness and intrusion level of these methods varies. Examples of these are honeypots, botnet takeouts, and entering the attackers' network to gain information or rescue stolen information.

Some measures that are used in defensive work (such as penetration testing within one's own network in order to patch the security breaches when found) can be used in another's network, making use of the same testing measures this time in an offensive cyberspace operation. It is also worthy of mention that the examples of AI-supported methods discussed by in [18] and labeled as offensive operations can also be classified as malicious use of AI, as in [2].

3 Background: Ethical considerations of AI usage

Concern for recognizing and adhering to relevant ethical norms and standards is often voiced by national and military leaders. The U.S. Department of Defense "Joint Artificial Intelligence Center" (JAIC), established in 2018, for example, lists leadership in "military ethics and AI safety" as among its "five pillars of AI strategy" [20]. Interestingly, neither the above-mentioned ENISA AI cybersecurity challenges report [2] nor NATO's cyberspace operation doctrine [4] includes explicit ethical questions or challenges among the otherwise vital issues targeted for closer research. Although much has now been written on the ethics of cyber operations generally, the ethics of AI use itself in cyber operations to date has largely gone unaddressed. What exactly such vague and generalized expressions of concern mean in practice is thus far from clear. And efforts to "operationalize" or otherwise use "ethical norms" in the development and use of AI-enhanced military technologies generally are not obviously apparent.

One recent study of ethics aimed at both AI lifecycle actors and end users, "Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications" [21] enumerates normative implications of existing AI ethics guidelines for developers and organizational users of AI generally. They describe eleven principles, which comprises 91 guidelines, most of which (82) are drawn from a general ethics overview by Jobin et al. [22].

1. **Transparency** in all facets of AI technology (data, algorithms and decision making) throughout the phases of development and use.
2. **Justice and fairness** guaranteed by developers and users so that AI does not discriminate against any groups or provide unfair outcomes.

3. **Non-maleficence** to avoid AI harming human beings.
4. **Responsibility** for AI actions and their consequences must always be traceable to a legal person.
5. **Privacy** should be ensured by AI-supported tools; the GDPR should apply.
6. **Beneficence.** AI use should be beneficial for humans.
7. **Freedom and autonomy.** AI use should strengthen democratic values and personal self-determination.
8. **Trust.** Developers and users of AI should demonstrate their trustworthiness, and the reliability of their AI systems.
9. **Sustainability.** Development and use of AI technology should be sustainable.
10. **Dignity.** AI use should not violate fundamental human rights.
11. **Solidarity.** AI use should promote social welfare and security.

Here we focus our ethical perspective on the end users of AI-supported cyberspace operations. Each of the effects of cyberspace operations (shown in Sect. 2.2) pose delicate, and sometimes non-apparent ethical and legal dilemmas. Despite AI's comparative advantages in pursuing these objectives, its utilization as a tool within cyber operations has exacerbated three problems: (i) inadvertent *escalation*, insofar as AI-enabled autonomous interactions without "humans in the loop" are resistant to normal measures of supervision and control; (ii) *proliferation*, insofar as AI reduces the human personnel needed for cyber operations, thereby decreasing the cost of these operations; and (iii) these lowered costs, in turn, permit more actors (state and non-state) to acquire the capacity to engage in cyber operations, and with that consequent multiplication of actors, the *attribution problem* grows proportionately more difficult.

Here the term "proliferation" (see ii above) is invoked in the sense that arms control analysts use the term, i.e., the spread of a deadly weapon. An offensive cyber capability can cause severe damage. Until recently only certain states that possessed large and dedicated cyber programs, with many personnel working together, were able to launch effective cyberattacks. But by using AI, smaller actors can enter the game. An example of this cyber proliferation (although not specifically with AI) is the spate of ransom incidents whereby municipalities and businesses have been locked out of their computer systems. The program being used for this purpose, "Wanna Cry," is reported to have been developed by the National Security Administration [23], and was somehow exfiltrated and released on the web (perhaps by a disgruntled employee) [24]. There it was obtained and used by cyber criminals, until it was detected and removed. This weapon has unfortunately proliferated. Our concern is that AI usage may only exacerbate this problem.

We do not mean to imply that current and future uses of AI have only ethical downsides. One obvious upside of AI is to strengthen defensive cyber capabilities via earlier detection and appropriate defensive response while easing the strain and manual workload of the human defenders. Another ethical upside in the war context is the enhanced ability to bring force to bear on an adversary with less destructiveness. One party might opt to shut down rather than physically destroy an adversary's power grid, for example, allowing services to resume when the conflict has been resolved.

3.1 GDPR for cybersecurity AI included

Before examining the ethics of AI from the end-user perspective (Sect. 4), we need to discuss the ramifications for AI-enhanced cyber operations of the General Data Protection Regulation (GDPR). GDPR is among the most stringent privacy and security laws in the world at present, and although it was drafted and passed by the European Union, it imposes legal and ethical obligations on organizations and states throughout the world who might be involved in targeting the security of, or collecting data about, citizens in EU countries [25]. This broad discussion, in turn, provides the background for consideration of this regulation's impact on the specific, case-by-case activities of end-users.

Cutting-edge cybersecurity systems will be increasingly data-driven [26]. As such, they will also risk running afoul of data protection regimes, such as the GDPR, as well as other laws restricting the use of AI. Ethical deployment of such systems thus should include some consultation with competent professionals capable of mapping the complex domain of requirements and exceptions characteristic of this body of law. This section will touch on basics, noting first the fundamental divide between state and non-state actions in this realm.

The GDPR applies to all companies, organizations, and government agencies that

> process [personally identifiable information] on individuals residing in the European Union ... regardless of where the entity is located [27].

State actors involved in the realm of policing enjoy great latitude under the GDPR [28]. When such actors collect data to prevent crimes or threats to public safety, the GDPR does not restrict their activities ([29], Art. 23). To the extent that defensive military cybersecurity is akin to crime prevention (prevention of harm to citizenry), the armed forces enjoy similar freedom from GDPR privacy restrictions. There are also specific derogations that Member States can apply in areas such as security and defense [28].

Cybersecurity reporting requirements may also fall under GDPR exemptions or derogations. For example, in some situations Member States require a wide range of entities to distribute information to other entities. Those entities required to engage in such information distribution remain exempt from the requirements of the GDPR so long as sufficient safeguards for the data are in place.

Moreover, Article 23 of the GDPR authorizes a Member State to, "when necessary and proportionate," restrict the scope of the obligations under the GDPR ([29], Article 23). Nevertheless, private entities developing cybersecurity systems should be mindful of GDPR protections. As U.S. Department of Defense attorney (and Professorial Lecturer in Law at George Washington University Law School) Brandon W. Jackson [30] has observed:

> Autonomous cybersecurity systems are driven by data, and the European Union General Data Protection Regulation (GDPR) is an unavoidable moderator in this regard. The GDPR places significant restraints on the collection and use of data in Europe. Moreover, the extraterritorial nature of the regulation compounds the impact it has on global industries.

While Jackson concludes that ”today’s AI-based cybersecurity systems are likely capable of complying with the GDPR,” he also cautions that ”absent a technical solution, maintaining compliance will becoming increasingly difficult as these systems achieve greater autonomy” [30] . These challenges will become more salient as AI components of cybersecurity systems become more important in the private sector.

4 Cyberspace operations and AI ethics guidelines for End-Users

In general, the ethical principles we recognize are easier to apply when our moral situation is relatively uncomplicated. Living by them will be challenged when we face situations of uncertainty, ambiguity, or great complexity. It is important to identify in advance some of the likely situations in which ethical dilemmas might surface and prepare ourselves through advance discussion of representative cases. When, for example, might the ethical principles that we normally follow need to be altered or even set aside? We may understand how moral rules apply in peacetime, but what do we do when we find ourselves in a complex and wholly unfamiliar crisis, such as war or equivalent forms of armed conflict?

In Sect. 2, we purposely omitted discussion of who might be conducting cyberspace operations, as these assignments vary depending upon national context. However, from the ethical point of view, we argue that all who are engaged in defensive cyber operations should also consider the ethical dilemmas that can arise in offensive cyber operations. As noted above, the roles, authorities, and responsibilities might change based on the situation, especially during a transition from peacetime to war. A civilian defensive cyberspace operator, for example, might be recruited by the military into an offensive cyber operation, much as during World War II civilians were called to duty as code breakers or to calculate bombing trajectories. To simplify and for purposes of the discussion that follows, we presume the ”end-user” to be a *military* cyber operator, as offensive cyber operations are usually carried out by defense forces. As the cyber operator inherits an occupation, we consider the two forms of occupational ethics that apply in this domain: military ethics and engineering ethics.

Transparency in AI-supported COs. Transparency can be understood in several different ways. Rather than attempt to offer an exhaustive catalog, it might help to think of undesirable and desirable transparencies, also including platforms other than cyber and AI-supported CO. In general, transparency in any domain and with respect to any weapon system or supporting technology or doctrine is something policy makers and their militaries are well advised to avoid. A similar logic applies to cyberspace operations, along with the specific tools and procedures that might be utilized in carrying them out [whether AI-supported or not], as these must remain organizational secrets.

Undesirable transparency comes in several forms, the most important of which in our context being the undesirable revelation of information that should be kept out of the wrong hands. As law professor David Pozen [31] has argued,

> As public institutions became subject to more and more policies of openness and accountability, demands for transparency became more and more threatening to the functioning and legitimacy of those institutions.

An even more direct threat is a real or potential adversary's ability to discern enough about a Blue team's (the term used for the "good" side in cyber exercises) capabilities, training, and doctrines, alongside intentions to neutralize some or all advantages those would confer. For examples, if the opposing Red team knows Blue's catalog of zero-day exploits useful for disabling currently-fielded AI, and also knows under what circumstances each zero-day exploit would be deployed and therefore revealed, Red team has an advantage.

In both offensive and defensive operations, the capabilities of AI-supported COs are rarely revealed intentionally. Occasionally, a nation might decide to send a message of national will or technological superiority in order to control Red team's psychology and motivate them to act in certain ways. But this is relatively rare.

The plan for a cyber operation must be issued in advance and approved by the relevant authority, and this requires some level of transparency. The plan should include a description of the rules of engagement (ROE), as well as define the standing authority and expected effects of COs [32]. The plan should also comply with applicable international law. Here the two *Tallinn manuals* provide guidance to end-users in determining when, where, and how compliance with relevant international law impacts their anticipated operations. (e.g., *Tallinn Manual 2.0* [14]: Rule 103).

Transparency has another meaning in the realm of development as well as in testing, evaluation, verification, and validation (TEVV) [33]. The essence of transparency here is that governments and their militaries should acquire and field only technologies with known or at least knowable effects under given sets of conditions. Similarly, industries and other organizations that develop systems which then pass TEVV regimens should emerge as open books; there should be no surprises once the systems are fielded. This sense of transparency could require a degree of flexibility with respect to strong AI.

It is conceivable that some aspects of a system's or weapon's transitions through "sense-think-act" iterations will not be fully understood, even when the TEVV process reveals regularity sufficient to engender the confidence a nation needs to field the system. "Explainable AI," however, helps increase transparency between AI and its users. It incorporates four principles [34]: explanation, meaning, accuracy, and limits of applicable knowledge. By following these principles in designing AI-systems, the point is to ensure that they operate only within the context for which they are designed such that the outputs are accompanied with reasoning that is understandable to the different user groups.

Desirable transparency can also be presented to all sides as a shared acknowledgment of unsafe practices that all should avoid for their mutual benefit. As with nuclear weapons, states have agreed on some desirable safety practices. Work is

ongoing to establish similar practices also for use of AI (e.g. the Global partnership of artificial intelligence [35]).

Thus we have two distinct kinds of transparency: one that pertains to weapon developers and the other that pertains to end users in the cyber battlespace. The first (developers) must demonstrate (be transparent about) the efficacy and safety of the systems they have designed. The second group (cyber operators) ordinarily aim to surprise; in this sense transparency (vis-à-vis the enemy) is inimical to their mission. Sometimes, however, cyber weapons are put to deterrent use; in this instance, a military will deliberately reveal details about its capabilities to dissuade an enemy from some line of action. In this context, transparency will be militarily appropriate. Moreover, apart from transparency toward enemies, military personnel who use AI-based cyber weaponry must, on demand, reveal to their superiors and others engaged in post-battle assessments what exactly happened when these weapons were deployed. In this sense, a strict obligation of transparency exists. To sum up, the transparency appropriate to the battlespace falls within the purview first and foremost of military ethics, while the transparency incumbent on developers mainly pertains to engineering ethics. There are evidently areas of overlap; for instance, engineers have an ethical mandate, clearly recommended by their code of ethics, not to field weapon systems incorporating military decision-algorithms (such as the US Air Force's Arachnid) unless the TEVV process has ensured there will be no surprises once the AI-supported technologies are finally deployed. Cyber operators, likewise, must acquire a basic understanding of the safety constraints of the systems they use, constraints that pertain chiefly to engineering ethics.

There should be something less than full transparency between state-security organizations that conduct cyberspace operations and the general public. The same applies for the individual users of the AI. Trade and national security secrets (particularly offensive cyber capabilities), must never be disclosed to friends, family, or the public. In another sense, however, the engineering-based transparency between the AI-supported tool developer and the end-users requires that the staff of the client security agency or organization understand how the AI technology they are using really works, such that their decision-making about its employment is fully informed. In this instance transparency is desirable and even required.

Justice and fairness in AI-supported COs. Concerns regarding justice and fairness often focus on implicit bias in databases that might result in unjust discrimination in operational outcomes. AI development depends on having sufficient data necessary to achieve optimal performance. We know that biased data sets can create problems in AI algorithms: e.g., the well-known discovery that some facial-recognition programs have failed to recognize Black women more often than they have failed to recognize White men. This phenomenon has obvious ethical implications. One solution is to increase the amount of data available to the AI as it learns to do its tasks. However, the quest to acquire the greatest possible volume of data can threaten individual privacy and corporate security.

Furthermore, while the problem of bias in data sets and among designers, testers, and operators is important, the general concerns for justice and fairness in a military

setting might encompass even more. The sorts of calculations implied by the *in bello* principle of proportionality, and the kind of deontic work that underlies the principle of discrimination, require an underlying conception of justice. Lacking such a conception, how would one know what to count as good or bad in a proportionality calculation, and how would one judge the aptness of rules for determining who is a combatant and who not as prerequisite to applying the principle of discrimination?

Consider the following three situations in which an AI-enhanced CO might introduce more complications than the human-controlled CO is causing today. The first one is the pace of move-countermove. Already today, the need for countering a cyberattack in cyberspace can be too time-limited to leave room for ethical considerations. Therefore, these considerations should be discussed and established within the ROEs beforehand, prior to any conduct of specific cyber operations. This requirement is also specified in NATO's cyberspace operation doctrine. Once AI-enhanced tools are introduced, however, cyber-cyber interactions can take place even faster, affording little or no chance of detecting errors before disaster occurs, resulting in an outcome ranging from an error in discrimination to a grave war crime.

A second situation arises from our basic ignorance of why attacks and counter-attacks unfold as they do. Suppose that an accidental but destructive attack occurs using AI-supported CO. It is hard, if not impossible, to reverse engineer from effect to algorithmic cause [36]. It is one thing to ask for lenience at the state level while providing a clear explanation of what transpired, and quite another only to be able to state: "We didn't intend for those bad things to happen, and we don't know how they happened."

The third situation arises from the ease of introducing multi-dimensional distancing. Much has been written about the moral dangers of "standoff" weapons. The foundational concern is that as the spatial or temporal distance between a would-be shooter and her intended victim increases, the shooter is less likely to suffer the visceral horror that would discourage her from taking another human life. To take just one possible example, an AI-supported CO can be psychologically tailored to avoid the qualms of legislatures, command authorities, and operators, and thereby encourage approving or doing violence that otherwise would not be countenanced. In the cyber domain, and especially with the aid of AI, tailoring a weapon to combine different types of distancing would be relatively simple.

Non-maleficence in AI-supported COs. It is tempting to say that with respect to non-maleficence and benevolence, AI-supported cyber defense and offense are not significantly different than operations in other domains fought with non-cyber means. Indeed, there are many enlightening similarities and analogies among domains and weapons.

The *Tallinn Manual (2.0)* [14], for example, states that similar to other dual-use objects, "Cyber infrastructure used for both civilian and military purposes is a military objective" (Rule 101) and therefore constitutes a legitimate military target. The manual, however, forbids the use of cyber "booby traps" associated with objects specified in the law of armed conflict (e.g., medical help) (Rule 106), giving a

specific example of malware embedded in fake emails from medical personnel causing physical illness (Rule 106, Explanation 4).

Like traditional warfare, offensive cyber operations can be used to harm civilian noncombatants: indirectly, by affecting the systems they use, but also directly (for example through malware embedded in digital medical devices). Collateral damage is also possible with offensive CO due to the connectivity of different systems and networks. Dual-use targets are likewise an issue of concern in cyber operations just as in conventional or kinetic operations. For example, disrupting traffic in networks which might be used for both military and civilian purposes will have an impact on civilian operations, and not just on the military.

But there is one curious pattern in cyber operations that is closer to ancient than modern 3-D warfare, a pattern one might describe in terms of reusability. In ancient times, many if not most weapons were ”recyclable.” Thus, if a Greek hoplite dropped his dory, an enemy soldier could pick it up and use it, perhaps after modification. Even some arrows (or at least their heads) could be recycled. In each of these cases, the artifact that does direct harm in war can be reused if captured or found. At a certain point in history, however, as weapons technology (and perhaps a facet of human psychology) made standoff combat preferable, most of war's artifacts of direct harm could no longer be recycled. By the twentieth century, the most traumatic killing was done by projectiles or other means - bombs and missiles and shells and bullets and poison gasses - that could not be recovered and reused. (The means of delivery of these artifacts of direct harm could be reused, but not the things that penetrated flesh and bone or respiratory systems.)

Many defensive and offensive cyber weapons at present are conceivably reusable if stolen or recovered after they are ”fired,” with the caveat that many are context-specific. The Stuxnet/Olympic Games code was useful to penetrate an industrial control system fronting centrifuges and creating a man-in-the-middle deception. Defensive measures such as coded encryption algorithms are likewise reusable in certain contexts. The ”WannaCry” software weapon cited above constitutes another example. Aside from its use after having been stolen, it is unclear whether the same weapon, if it had been used by the US first, would have been recoverable, and if so, whether it could subsequently have proven useful in attacking US interests (a prospect that should motivate us to reflect further on symmetry).

The point is that once a nation develops an offensive cyber weapon, it is important that a blend of pessimism and humility motivate it to build defenses against any capture or reuse of its own innovation. And if this proves to be impossible, it must seriously consider whether the weapon should be built at all. Similarly, once a nation finds a potent defensive cyber tool, it behooves it to envision possible offensive means to overcome it. In general, a nation cannot be sure its offensive and defensive weapons will not be stolen [24] or otherwise turned against it by malicious users. In this sense, cyber weapons, though standoff, have conceivably returned us to a problem last prevalent in ancient combat.

From the non-maleficence and benevolence perspective, we need to accept the fact that cyber weapons will often ”leak” beyond their intended targets (as Stuxnet did), and that they can be recovered and possibly be re-used or reverse-engineered

for unintended malevolent purposes. Therefore, there is a need to think hard about proportionality and discrimination not only with respect to the weapons' effects on our adversary but also on ourselves and our allies (Stuxnet, for example, was famously designed to limit the useful prospects of capture or reuse ([37] p.58-60)). Similarly, when developing and fielding cyber defenses, we must think of our activities not only in terms of benevolent results on our own and our allies' behalf but also on the harm they might cause if the adversary appropriates them through theft or inference, aided perhaps by observing the effects of multiple, probing attacks.

What does this mean for the end-user? Based on NATO's cyberspace operation doctrine, offensive operations are not conducted without a decision from the highest decision level. It is ordinarily teams of cyberspace operators, however, that will in the end carry out the commands. They are therefore in a position not unlike platoons of conventional combatants, concerned with determining when it is ethically acceptable to attack or destroy an adversary. Other issues discussed above likewise fall finally to individual cyber operators (usually working in small teams), such as the responsibility for carefully developing, deploying and storing cyber weapons. The end-user at the tactical level can also serve as a knowledgeable advisor to command, offering input on the risks, as well as benefits gained from using specific cyber tools. All these factors conspire to place a burden of responsibility on cyber end-users (just as conventional combatants) to be knowledgeable about the ethical dilemmas attendant upon their own actions in carrying out the orders they might receive.

Responsibility in AI-supported COs. In the military setting, all missions have an assigned leader and therefore the ultimate responsibility for effects of specific CO's utilized falls by law to the mission commander. *Tallinn Manual 2.0* [14], Rule 85, specifically states the following: a) Commanders and other superiors are criminally responsible for ordering cyber operations that constitute war crimes; b) Commanders are also criminally responsible if they knew or, owing to the circumstances at the time, should have known their subordinates were committing, were about to commit, or had committed war crimes and failed to take all reasonable and available measures to prevent their commission or to punish those responsible.

To avoid criminal activities, NATO's doctrine specifically states that, prior to an offensive cyberspace operation being carried out, the discussion at higher decision levels, including legal experts, must have taken place. Thus, before the end-user proceeds to undertake an action, the decisions to carry out the specified action have been reached higher up in the appropriate chain of command.

However, accidents may yet occur and ignorance of relevant software design or operational procedures can still bring about the unfortunate result that the end-users actually conducting the cyber operations are discovered to have acted wrongly. In these cases, it is for digital forensic investigation to bring forth the evidence, and of the judicial system to use that evidence to prove guilt [38].

But what happens if or when an AI-enhanced system itself haphazardly runs amok? Absent a thoughtful advance notion of collective or systemic responsibility (and liability), such accountability is likely to be assigned arbitrarily or by rote (as described above) to the ranking member military unit directly implicated in

the damage caused by AI. It might not make sense, let alone seem altogether just, however, simply to assign blame to the ranking officer or commander when AI-enhanced cyber activities go awry. But would not the threat of doing so, at least, serve as a catalyst to crack open any ”black boxes” and demand AI systems that are fully explainable, knowing in advance who otherwise will ultimately be ”on the hook” in the worst case?

An additional wrinkle regarding responsibility is compartmentalization. A familiar bureaucratic principle in defense and security operations requires, for example, that one have not only the appropriate clearance to be read in on any given highly-classified topic, but also the demonstrable need to know. The dual requirement leads naturally to the compartmentalization that ”stovepipes” whole communities in order to keep the information they need more secure than it would otherwise be. Compartmentalization increases the risk of accidents, as linkages between parts of an interlocking system are poorly understood by its individual operators, who focus solely on their specific tasks, thereby ignoring the ramifications for the overall system (see [39], p. 103-167), on a friendly-fire incident that took place in Northern Iraq in 1994).

A leader at any level of the command hierarchy will probably be too busy (and perhaps also lack the specific expertise) to evaluate the ethical aspects of any complex weapon system, cyber or not, such that assigning responsibility for technical failures to that leader would be somewhat arbitrary in any case. But if one adds the black-box aspect of AI and the compartmentalization that is common in the entire cyber operations realm, individual leadership responsibility for tech-based failures of discrimination or proportionality will seem increasingly far-fetched. The inevitable conclusion is that ”responsibility” will often be systemic in nature, rather than traceable to specific individuals (see [39]).

Responsibility for deploying and using AI that subsequently runs amok is not crystal clear and will most likely vary dramatically from case to case. What does this factor mean for the end-user? This likely transcends military ethics questions related to responsibility in operations, but also invokes purely personal values. How willing is any given operator to utilize a tool whose functionality he or she is unsure of? How much effort will the individual operator expend to learn and understand the tool she is using? How much autonomy is the individual operator willing to delegate to an AI-enhanced cyber weapon or system itself, if it is they (rather than it) who will nonetheless be held responsible in the end for its proper functionality? Yet again, how willing is the individual operator likely to be to confess his or her own mistakes, or even to report incriminating activity by others, in cases where that operator has either authorized or used AI-supported tools even before official permission has been given, or in contexts where the AI-supported tools were not intended to be used?

Privacy in AI-supported COs. When conducting defensive and offensive COs, the full privacy of the user cannot be given. Both defensive and offensive CO employ AI-enhanced tools for network and system monitoring and analysis; however, the sensitivity and level of detail vary. AI-supported tools will obviously make the monitoring and analysis of information faster. The problem with guaranteeing

individual identity or privacy during such operations is that everything will be stored, thereby offering the possibility of use of stored data for illegal purposes.

Similarly, as discussed under the category of transparency above, private information that a nation has stored on its own equipment can be used against it by any adversary if that information is exfiltrated or stolen. Such data can also be misused by one's own operators when, for example, while monitoring the networks that one is assigned to protect also affords a possibility for spying on one's own colleagues. It is therefore both an operational and engineering-related question to determine what kind of information is relevant to monitor, store and analyze in each operation.

Another hypothetical ethical dilemma arises when ongoing criminal activity is inadvertently discovered, but when reporting or acting on this discovery might also compromise the primary mission. For example, when spying on adversaries' networks, one might notice ongoing criminal activities (e.g., child pornography). What would the appropriate individual choice or organizational ethical decision be: to save a child but "blow the cover," or decide instead to ignore the discovery and continue the operation? Which good is the "greater good?"

This raises the additional specter of what has come to be called "lawfare" i.e., using provisions of relevant law as weapons against an adversary [40]. For example, the *New York Times* reported that the SolarWinds attackers had used US-based servers to stage their efforts in order to avoid NSA scrutiny, which would in that case be prohibited by statute 215 of the Patriot Act from surveillance of American citizens [41]. Why wouldn't a hostile actor twist any tool at hand, including the Fourth Amendment guaranteeing freedom of expression, into a useful weapon in the cyber arena?

Intelligence expert Jim Baker [42] worries about the vast amounts of information concerning the behavior of US citizens that can be gleaned from emerging 5G networks, an ever-enlarging IoT, and other sources. Hostile powers could utilize AI to mine such data to gain the ability to understand, then predict, and finally manipulate behaviors in the US to suit their ends. Presumably a primitive version of this tactic was utilized in the Russian disinformation and voter manipulation efforts during the 2016 US elections. Baker concludes, however, by pointing out that US counterintelligence will be an obvious target: ominously, undermining the efforts of the security "guardians" by understanding them, predicting their next moves, and finally manipulating them through information operations.

Even though such information operations themselves are not part of this chapter, the worries outlined by Baker are relevant for end-users to understand. Variation in internal laws and compliance of international rules between nations differentiates possibilities of methods used in defensive and offensive COs. GDPR mentioned in Sect. 3.1 is an example, where there are major differences in compliance required between nations. A cyber incident or a conflict (or even war) is not considered fully fair when opposing sides are constrained by law in different ways. Those final examples point in turn to the characters of the individual operators. How willing is an end-user to adhere to legal restrictions applying to her or to the methods she uses, if the adversary is not likewise bound by them?

Beneficence in AI-supported COs. AI use should, on balance, promise at least to prove beneficial for individual well-being. It should be used for securing the common good, social good and peace. Here, in general, the ethical dilemma is to define whose ”well-being” warrants consideration. In cyberspace operations, there is always ”us” and ”them”: the Blue team versus the Red team, the ”nation” and its adversaries and competitors. But who finally is ”us” and who constitutes ”them”? An individual user is likely to find herself belonging to several different legitimate groups of stakeholders. She will be part of an organization, such as a cybersecurity unit, to be sure. But there might be internal conflict inside the organization itself dividing it into separate, competing teams, each of which also includes other individuals with disparate goals, each one worrying ”what does this mean for me?” Also, an organization's goals might sometimes fail to align with its supervening nation's goals. Whose overall welfare then has the highest priority? And even more to the point: which standpoint enjoys the higher priority, the individual operator's own nation's welfare or interest, or the standpoint of international humanitarian laws and rights of others, even including the enemy?

The use of AI decreases the manual workload in cyberspace operations, for instance, by analyzing net traffic and disclosing anomalies or scanning for system weaknesses, all of which benefit the end-user. AI can also provide a vastly enhanced background for decision-making based on predictions (for example mapping possible collateral damage), allowing cyber operators to discern which offensive cyber operations would lessen the harm done to the enemy, but still prove of maximum benefit for ”us”. In this fashion, AI allows concern for beneficence to be looked at comparatively with reduction of maleficence to adversaries, thereby significantly improving compliance with the requirement of proportionality of means and ends during conflict.

Freedom and autonomy in AI-supported COs. China's current efforts aimed at finding ethnic minorities and categorizing them as potential threats [43] is an example of a large-scale use of AI-supported tools in cyber domain in which individual freedom is severely threatened. Methods and tools used in the cyber operations discussed in this chapter could, in theory, likewise be used in finding and categorizing people according to some threat assessment. However, the regulations and laws stemming from the GDPR set well-defined limits to the use of data collection for such purposes. First the persons whose data would be collected must give their prior, informed consent. Second, access rights to the data are also limited - meaning, for example, that military units do not automatically have authorized access to the personal finance data or phone traffic data of civilians. Likewise, police units do not have access to individual health records without either prior permission or legal warrants.

Surveillance can also be carried out within an organization's own network, however, as well as blocking access to information flow. In some organizations, for example, social media sites are blocked during working hours via organizational tools whose use is perfectly legal. Similarly, as discussed with respect to terms of individual privacy, employee preferences can be catalogued and may even be

used against them inside an organization. Hence, the risk of discrimination and of overriding individual freedom and autonomy exists internally.

For the end-user perspective, two familiar ethical questions can arise. One is the same as in privacy discussions generally: namely, whether to use the confidential information of one's colleagues for one's own purposes. The second question is, what does working under the prospect of such constant surveillance do to the individuals surveilled?

Trust in AI-supported COs. Trust is closely related to the first principle, transparency, inasmuch as both have individual, organizational and state levels. Trust develops over time, and that applies to both human relationships and relationships between humans and their technologies. Consider the "smart phone." We use it for a variety of purposes and our trust in its safety has been established. We are aware of the downsides (for example, that it is constantly collecting data on us) but its utility in enabling our manifold daily routines and especially its instant connectability simply outweigh these well-known downsides.

Transparency regarding how AI functions is one of the key points for maintaining public trust (as noted above). However, there are other trust-related issues in using AI tools that pose ethical questions for end-users to consider. One such question, also related to transparency, is the degree to which end-users credit the results given by AI tools. One might say that the higher the level of engineering education the user has, the more the consequent trust in the results of AI-assisted operations can be grounded confidently in knowledge, rather than in mere faith. Perhaps a person possessing a high degree of knowledge is better suited than others, for example, to challenge the results and decisions given by AI.

Being a trustworthy person is also often cited as a chief virtue of a good person. What does trustworthiness mean in cyber operations? Is it a loyalty to your unit, to your people, to a state, or to humankind? Here there are once again several layers of significance to ponder, but at least two are straightforward to identify: viz., being a whistle blower, and being an insider threat. Where does one draw the line between staying loyal and remaining silent instead of becoming a whistle blower by bringing attention to ongoing misuse of a cyber tool? Or when does one decide to become an insider threat to one's organization or government, either to exact personal revenge, or from having been influenced by others?

Sustainability, Dignity and Solidarity in AI-supported COs. Obviously, cyber operations need energy. However, it is difficult to generalize whether energy use is less or more with AI-supported tools than without. Nonetheless, sustainability from the energy perspective is something that developers of AI tools should factor into the design of their systems. Offensive cyberspace operations, however, can have environmental effects. For example, disturbing the functionality of a dam can cause a flood. However, the use of cyberattacks for delivering disturbing effects need not destroy the system itself; the goal can rather be to disable it for a particular time. In this sense, cyberattacks are less harmful to society than kinetic operations (bombing raids, for example) and in this sense are more sustainable.

Dignity means respecting human rights and recognizing that each person has inherent value. In general, AI supported tools should be used in such a way that dignity is preserved. Similarly, use of AI should promote social security and cohesion, and not undermine solidarity. Cyberspace operations that we focus on in this chapter do not target specific individuals directly. Neither are they used to create, manipulate, and spread false information. Those are instead means that pertain to Information Operations. However, the effect on dignity of the cyberspace operations we consider in this chapter can be indirect. AI-supported tools can be used, for example, to extract sensitive information from a database owned by an organization, a company for instance, that is to be harmed. If the data is subsequently leaked, thereby damaging the reputation of the company, its customers, whose personal data has now been publicly released, are the surrogate victims, collateral damage, as it were.

Similarly, solidarity might be indirectly affected. For example, offensive cyber operations can be used to degrade, disrupt, and destroy supply chains of the enemy. Even if the direct effect is on those supply-chain systems, the indirect effect can fall upon civilian groups needing humanitarian help. Here again, the rules of engagement should also be discussed with and among the tactical-level end-users of the AI, both from the military-ethical point of view but also from a personal point of view. With which commands is the end-user ultimately willing to comply? What can AI decide alone, and where should a human be part of the process?

5 Conclusion

In this chapter, we have attempted to enumerate and describe many of the benefits, as well as some of the moral challenges, that end-users in AI-enhanced cyber operations, both offensive and defensive, are likely to face. Many of these challenges have to do with individual and collective accountability for any negative or unintended consequences of such operations, coupled with the thorny problem of transparency and attribution of those consequences. These dilemmas are particularly intractable when gauged asymmetrically: between operators and their agencies, who are tasked with promoting ethically responsible operations in cyber conflict, on the one hand, and those, on the other hand, who decline to become encumbered by any such scruples. It helps to restore balance between parties to cyber conflicts if the ethical challenges can be identified in advance, and morally responsible considerations ”baked into” the strategies formulated in response. Thereby, one avoids the need to introduce moral considerations ”on the fly” as additional constraints to be imposed upon decision-making and time-sensitive action in the midst of conflict. This essay intends to initiate, if nothing else, a serious discussion about the important task of anticipating and developing strategic responses to cyberattacks and intrusions; responses that are reasonably guaranteed, in themselves, to uphold the values of our respective nations and allies, even while we are engaged in the complex and time-sensitive tasks of providing security for our citizens' lives and property in the cyber domain.

References

1. K. Ayoub, K. Payne, Journal of Strategic Studies (2016)
2. ENISA ad hoc Working Group on Artificial Intelligence, *AI CYBERSECURITY CHALLENGES-Threat Landscape for Artificial Intelligence* (2020). URL https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges
3. EU High-Level Expert Group on Artificial Intelligence, *A Definition of AI: Main Capabilities and Disciplines* (2019). URL https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines
4. NATO, *Allied Joint Doctrine for Cyberspace Operations AJP-3.20* (2020)
5. N.J. Nilsson, *The Quest for Artificial Intelligence* (Cambridge University Press, 2010)
6. G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido, M. Marchetti, in *10th International Conference on Cyber Conflict (CyCon), Tallinn* (2018), pp. 371–390. DOI 10.23919/CYCON.2018.8405026.
7. D.S. Berman, A.L. Buczak, J.S. Chavis, C.L. Corbett, Information **10**(4), 122 (2019). DOI 10.3390/info10040122
8. A. Gilli, M. Gilli, A.S. Leonard, Z. Stanley-Lockman, "NATO-Mation": Strategies for Leading in the Age of Artificial Intelligence. NDC Research Paper 15 in NATO Defense College "NDC Research Papers Series" (2020)
9. L. Vaccaro, G. Sansonetti, A. Micarelli, Computers **10**(1:11) (2021). DOI 10.3390/computers10010011
10. A. Akram, J. Lowe-Power, arXiv (2020)
11. M. Comiter. Attacking Artificial Intelligence-AI's Security Vulnerability and What Policymakers Can Do About It. Belfer Center for Science and International Affairs (2019). URL https://www.belfercenter.org/publication/AttackingAI. Cited 3 May 2021
12. K. Hartmann, C. Steup, in *12th International Conference on Cyber Conflict (CyCon), Estonia* (2020), pp. 327–349. DOI 10.23919/CyCon49761.2020.9131724
13. Joint Artificial Intelligence Center (JAIC). Integrating AI and Cyber into the DoD (2019). URL https://www.ai.mil/blog.html. Cited 3 May 2021
14. M.N. Schmitt, L. Vihul (eds.), *Tallinn Manual 2.0 - On the international law applicable to cyber operations, 2nd Ed.* (Cambridge University Press, 2017)
15. M.N. Schmitt (ed.), *Tallinn Manual 1.0-On the international law applicable to cyber warfare* (Cambridge University Press, 2013)
16. *San Remo Manual on International Law Applicable to Armed Conflicts at Sea, 12 June 1994* (1994). URL https://ihl-databases.icrc.org/ihl/INTRO/560
17. T.C. Truong, Q.B. Diep, I. Zelinka, Symmetry **12**(3), 410 (2020). DOI 10.3390/sym12030410
18. N. Kaloudi, J. Li, ACM Computing Surveys **53**(1) (2020). DOI 10.1145/3372823
19. J. Pattison, European Journal of International Security **5**(2), 233 (2020). DOI 10.1017/eis.2020.6
20. Joint Artificial Intelligence Center (JAIC) (2018). URL https://www.ai.mil/about.html
21. M. Ryan, B.C. Stahl, Journal of Information, Communication and Ethics in Society (2020). DOI 10.1108/JICES-12-2019-0138/full/html
22. A. Jobin, M. Ienca, E. Vayena, Nature Machine Intelligence **1**(9), 389 (2019). DOI 0.1038/s42256-019-0088-2
23. N. Harley. North Korea behind WannaCry attack which crippled the NHS after stealing US cyber weapons, Microsoft chief claims. The Telegraph (2017). Cited 3 May 2021
24. B. Buchanan, *The Hacker and the State* (Cambridge, MA: Harvard University Press, 2020). DOI 10.4159/9780674246010-004
25. B. Wolford. What is GDPR, the EU's new data protection law? GDPR.EU (2019). URL https://gdpr.eu/what-is-gdpr/. Cited 26 May 2021
26. T. Kontzer. What Does the Near Future of Cyber Security Look Like? A Roomful of RSAC Attendees Considered That, and Here Are the Takeaways. RSA Conference Blog (2019). URL https://www.rsaconference.com/library/blog/what-does-the-near-future-of-cyber-security-look-like-a-roomful-of-rsac-attendees. Cited 25 May 2021

27. D. Kawamoto. Will GDPR Rules Impact States and Localities? GovernmentTechnology (2018). URL https://www.govtech.com/data/Will-GDPR-Rules-Impact-States-and-Localities.html. Cited 3 May 2021
28. Council on Foreign Relations. Ctrl + Shift + Delete: The GDPR's Influence on National Security Posture. Net Politics (2019). URL https://www.cfr.org/blog/gdpr-influence-national-security-posture
29. European Commission, *General Data Protection Regulation (GDPR)* (2018)
30. B.W. Jackson, Minnesota Journal of Law, Science & Technology 169 **21**(1) (2020)
31. D.E. Pozen, YALE LAW JOURNAL **128**, 100 (2018)
32. C.R. Kehler, H. Lin, M. Sulmeyer, Journal of Cybersecurity **3**(1) (2017). DOI 10.1093/cybsec/tyx003
33. MITRE, *Verification and Validation, Systems Engineering Guide* (2013). URL https://www.mitre.org/publications/systems-engineering-guide/se-lifecycle-building-blocks/test-and-evaluation/verification-and-validation
34. P.J. Phillips, C.A. Hahn, P.C. Fontana, D.A. Broniatowski, M.A. Przybock, *Four Principles of Explainable Artificial Intelligence, NISTIR 8312* (2020). DOI 10.6028/NIST.IR.8312-draft
35. Global partnership of artificial intelligence (GPAI). Working group on responsible AI (2020). URL https://gpai.ai/projects/responsible-ai/. Cited by 26 May 2021
36. W. Knight. The Dark Secret at the Heart of AI (2017). URL https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/. Cited by 8 May 2021
37. G. Lucas, *Ethics and Cyber Warfare: The Quest for Responsible Security in the Age of Digital Warfare* (New York: Oxford Univeristy Press, 2017)
38. R. Crootof, 164 University of Pennsylvania Law Review 1347 (2016)
39. N.G. Leveson, *Engineering a Safer World* (Cambridge, MA: MIT Press, 2011)
40. O.F. Kittrie, *Lawfare* (Oxford University Press, 2016)
41. D.E. Sanger, N. Perlroth, J.E. Barnes. As Understanding of Russian Hacking Grows, So Does Alarm, New York Times. New York Times (2021). URL https://www.nytimes.com/2021/01/02/us/politics/russian-hacking-government.html. Cited by 26 May 2021
42. J. Baker. Counterintelligence Implications of Artificial Intelligence - Part III (2018). URL https://www.lawfareblog.com/counterintelligence-implications-artificial-intelligence-part-iii. Cited by 8 May 2021
43. Human Rights Watch (HRW). China: Big Data Fuels Crackdown in Minority Region (2018). URL https://www.hrw.org/news/2018/02/26/china-big-data-fuels-crackdown-minority-region. Cited 8 May 2021