International Review of the Red Cross (2020), **102** (913), 495–509.
Digital technologies and war
doi:10.1017/S181638312000048X

IRRC

REPORTS AND DOCUMENTS

Reports

This section of the Review *provides a summary of new ICRC-affiliated reports relating to this issue's theme of "Digital Technologies and War", including the executive summaries of three such reports. For access to the full reports, please follow the links provided.*

: : : : : : :

Symposium Report: Digital Risks in Armed Conflicts, October 2019

This report summarizes the key findings and action points coming out of a symposium on digital risks in armed conflicts and other situations of violence, held in December 2018 by the International Committee of the Red Cross (ICRC). The two-day event brought together representatives from humanitarian organizations, academia, tech companies and governments, as well as donor representatives. Discussions focused on how the use of digital technologies, including by parties to conflict and private companies, but also by the humanitarian sector as part of a humanitarian response, could put crisis-affected people at risk and make them more vulnerable both on- and offline.

Available at: https://shop.icrc.org/symposium-report-digital-risks-in-armed-conflicts-print-en.

The Humanitarian Metadata Problem: "Doing No Harm" in the Digital Era, October 2018

New technologies continue to present great risks and opportunities for humanitarian action. To ensure that their use does not result in any harm, humanitarian organizations must develop and implement appropriate data protection standards, including robust risk assessments. However, this requires a good understanding of what these technologies are, what risks are associated with

their use, and how we can try to avoid or mitigate those risks. This joint report by Privacy International and the ICRC aims to provide people who work in the humanitarian sphere with the knowledge they need to understand the risks involved in the use of certain new technologies. The report also discusses the "do no harm" principle and how it applies in a digital environment.

Available at: www.icrc.org/en/download/file/85089/the_humanitarian_metadata_problem_-_icrc_and_privacy_international.pdf.

Handbook on Data Protection in Humanitarian Action, Second Edition, May 2020

This handbook was published as part of the Brussels Privacy Hub and ICRC's Data Protection in Humanitarian Action project. It is aimed at the staff of humanitarian organizations involved in processing personal data as part of humanitarian operations, particularly those in charge of advising on and applying data protection standards. The handbook builds on existing guidelines, working procedures and practices established in humanitarian action in the most volatile environments and for the benefit of the most vulnerable victims of humanitarian emergencies. It seeks to help humanitarian organizations comply with personal data protection standards, by raising awareness and providing specific guidance on the interpretation of data protection principles in the context of humanitarian action, particularly when new technologies are employed.

Available at: https://shop.icrc.org/handbook-on-data-protection-in-humanitarian-action-print-en.

The Potential Human Cost of Cyber Operations, May 2019

Executive summary

Cyber operations during armed conflicts: Assessing the challenges for international humanitarian law

The use of cyber operations during armed conflicts is a reality. While only a few States so far have publicly acknowledged that they use them, cyber operations are a known feature of present-day military operations and the use of them is likely to increase in the future.

This new reality has triggered a debate regarding the rules of international law that apply to such operations. In this debate, the ICRC has recalled that during armed conflict, cyber operations are subject to the rules of international humanitarian law (IHL).[1] It is nevertheless clear that cyberspace and these new

1 See, in particular, ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts*, Geneva, 2015, pp. 39–44. The restrictions imposed by IHL do not legitimize the use of force in cyberspace, which remains governed by the United Nations Charter.

https://doi.org/10.1017/S181638312000048X Published online by Cambridge University Press

military operations raise a number of questions as to precisely how certain rules of IHL – which were drafted primarily with the kinetic realm in mind – apply to cyber operations.

Assessing these questions requires an understanding of the expected use and military potential of cyber technology. What aims may belligerents want to achieve by using new tools at the strategic, operational or tactical levels during conflicts? How does this new technology compare to other, existing means of warfare?

Furthermore, to assess how IHL protects civilians in armed conflict, and whether further regulation is needed, lawyers and policy-makers require an understanding of the actual or potential human cost of cyber technologies. Indeed, one of the main aims of IHL is to protect civilians from the effects of military operations.

Purpose and scope of the meeting on IHL and cyber operations

As part of its mandate to work for the clarification of IHL and, if necessary, prepare any development thereof, the ICRC monitors the development of new technologies that are, or could be, used as means and methods of warfare during armed conflicts. This approach is based on legal, technical, military and humanitarian considerations, which are interrelated.

To develop a realistic assessment of cyber capabilities and their potential humanitarian consequences in light of their technical characteristics, the ICRC brought together scientific and cyber security experts from all over the world to share their knowledge about the technical possibilities, expected use and potential effects of cyber operations. The three-day meeting drew on the expertise of participants working for global IT companies, cyber threat intelligence companies, computer emergency response teams, a national cyber security agency, participants with expertise in cyber security (including that of hospitals, electrical grids and other services), participants with expertise in the development and use of military cyber operations, lawyers and academics.

States and militaries remain reluctant to disclose their cyber capabilities, including the details of cyber operations conducted in the context of armed conflicts, and little is known about the few acknowledged cases. Therefore, the experts discussed a number of the most sophisticated known cyber operations, regardless of whether they occurred in the context of an armed conflict or in peacetime. Examining the technical features of these attacks and the specific vulnerabilities of the respective targets provides a powerful evidence base for what is technically possible during armed conflict. The meeting focused in particular on the risk that cyber operations might cause death, injury or physical damage, affect the delivery of essential services to the population, or affect the reliability of internet services. It looked at the specific characteristics of cyber tools, how cyber threats have evolved, and the cyber security landscape.

Approaching the subject from a humanitarian law and humanitarian action perspective, the ICRC seeks a sober and – to the greatest extent possible – evidence-

https://doi.org/10.1017/S181638312000048X Published online by Cambridge University Press

based understanding of the risks of cyber attacks[2] for the civilian population. The meeting allowed the ICRC to confirm much of its own research, and to supplement it with highly valuable additional expert knowledge. The meeting was extremely useful in that it contributed to a nuanced picture of cyber operations, demystifying some of the assumptions that often surround discussions on cyber warfare.

Areas of concern

Discussions helped to put the spotlight on four areas of particular concern in terms of the potential human cost of cyber operations:

a. the specific vulnerabilities of certain types of infrastructure;
b. the risk of overreaction due to potential misunderstanding of the intended purpose of hostile cyber operations;
c. the unique manner in which cyber tools may proliferate;
d. the obstacles that the difficulty of attributing cyber attacks create for ensuring compliance with international law.

a. Specific vulnerabilities of certain types of infrastructure: Cyber attacks that may affect the delivery of health care, industrial control systems, or the reliability or availability of core internet services

Apart from causing substantial economic loss, cyber operations can harm infrastructure in at least two ways. First, they can affect the delivery of essential services to civilians, as has been shown with cyber attacks against electrical grids and the health-care sector. Second, they can cause physical damage, as was the case with the Stuxnet attack against a nuclear enrichment facility in Iran in 2010, and an attack on a German steel mill in 2014.

Cyber attacks that may affect the delivery of health care

The health-care sector is moving towards increased digitization and interconnectivity. For example, hospital medical devices are normally connected to the hospital's information technology (IT) system to enable automatic electronic filing. Connected biomedical devices, such as pacemakers and insulin pumps, make it possible to remotely monitor individual patients' health as well as the functioning of the medical devices themselves.

This increased digital dependency, combined with an increased 'attack surface', has not been matched by a corresponding improvement in cyber security. Consequently, this infrastructure is particularly vulnerable, with potentially serious consequences for health and life.

2 The terms "cyber attacks" and "cyber operations" are used throughout the report in a technical (mainstream or colloquial) sense and not as they may be understood under IHL, unless specifically stated.

https://doi.org/10.1017/S181638312000048X Published online by Cambridge University Press

Cyber attacks against industrial control systems, including those used in critical civilian infrastructure

Industrial control systems are protected by complex safety mechanisms and often have built-in redundancy to guarantee safety and reliability. For example, electrical networks are grids with multiple power sources to avoid widespread effects when one of their parts is affected. Nonetheless, attacks on specific nodes might still cause a significant impact, such as if a critical system (like a hospital) depends on a specific sub-system or node, or because they have cascading harmful consequences.

Carrying out a cyber attack against an industrial control system requires a certain expertise and sophistication, and often, custom-made malware. Such attacks have been less frequent so far than other types of cyber operations. Nonetheless, their frequency is reportedly increasing, and the severity of the threat has evolved more rapidly than was anticipated only a few years ago. There is a risk that tools developed by the best-resourced actors may be repurposed or purchased by other actors who lack the expertise required to develop them from scratch. Moreover, there is a possibility that a number of undetected actors are capable of attacking industrial control systems.

Cyber attacks that may affect the reliability or availability of internet services

Cyber attacks that disrupt core internet services – such as the domain name system (DNS), which supports communications on the Internet – or disrupt the functioning of major cloud services may impact all services that rely on them. However, the risk of seriously compromising these core internet services was assessed by the experts as unlikely at the present moment thanks to the high degree of redundancy in the DNS and because major cloud providers tend to offer high security standards. If, however, such disruption were to occur, it could have widespread and potentially serious consequences, for example when life-saving services such as ambulances rely on the cloud.

Finally, distributed denial-of-service (DDoS) attacks have been used against services provided by governments for the population. Such attacks are carried out through increasingly large botnets. The arrival of the "Internet of things" will further increase the number of connected devices that could be used in such attacks. Furthermore, DDoS attacks might have a wider impact than expected by their author, in particular when information about the targeted network is incomplete.

b. Risk of overreaction due to the potential misunderstanding of the intended purpose of hostile cyber operations

Cyber operations can be broadly divided into two categories, depending on their purpose:

- activity encompassing reconnaissance, surveillance and the exfiltration of data and information, for example for espionage, often referred to as computer network exploitation (CNE), or "access operations";

https://doi.org/10.1017/S181638312000048X Published online by Cambridge University Press

- activity aimed at generating effects on a targeted system or device, such as tampering with data integrity (deletion, modification), affecting availability (disabling, including for prolonged periods of time), or causing physical effects, such as damaging the system, often referred to as a computer network attack (CNA), or "effects operations".

The distinction is primarily one of purpose. From a technical perspective, the initial steps of a CNE and a CNA to gain and maintain persistent access to the target may be identical. CNEs can then be turned into CNAs relatively simply, mostly through the use of specific payloads of a different nature. While the initial steps of the attacks may be tracked, it is often difficult to fully assess the attacker's purpose until the effect on the end target is actually achieved.

When the target does not know the actual purpose of the operation, its reaction may be to consider the potential worst-case impact that the attacker could achieve though a CNA and react in a stronger manner than it would have if it had known that the intended purpose of the attack was CNE. This escalation risk factor may give rise to a potentially harmful overreaction.

c. *Proliferation of cyber tools*

A third concern is the proliferation of cyber tools – an issue that in some respects raises concerns similar to those that may exist with regard to weapons proliferation or the proliferation of dual-use technology, although the specific nature of cyber tools must be taken into account.

Cyber tools and methods can proliferate in a unique manner that is difficult to control. First, cyberspace is a global domain: provided that the attacker can overcome the cyber security and defence measures in place, any network node and information residing on the network can be accessed from anywhere in the world. At the same time, cyber tools can be repurposed or re-engineered. The combination of these two characteristics means that when cyber tools are used, stolen, leaked or otherwise become available, actors other than those who developed them might be able to find them, reverse engineer them, and reuse them for their own purposes.

Finally, the fact that cyber tools and methods can be repurposed and reused is one of the factors making rapid and reliable technical attribution of cyber attacks a challenging process.

d. *Attribution of attacks*

While not a primary focus of the meeting, the discussions also touched upon the anonymity of attacks and the difficulty of attributing them to a specific actor, which is a fourth area of concern.

Cyberspace is a complex domain where multiple actors operate: individual hackers; criminal groups, potentially motivated by financial gain; States; non-State armed groups; and other non-State actors. Actors may also cooperate: for

https://doi.org/10.1017/S181638312000048X Published online by Cambridge University Press

example, States may buy cyber tools or have an operation performed on their behalf against a target they have identified.

Digital forensics and the capabilities of attribution of malicious cyber activity appear to be improving. Nonetheless, the ability of threat actors to obscure or effectively hide the origin of their operations on the Internet, compounded by the ability to buy, repurpose or re-engineer cyber tools developed or used by other actors, continues to make it difficult to rapidly and reliably attribute cyber attacks to a specific actor. This hampers the possibility of identifying actors who violate IHL in cyberspace and holding them responsible. This is a concern because to hold such actors responsible is one way to ensure compliance with IHL. It may also lower the threshold of using cyber attacks and of using them in violation of international law, because attackers can deny responsibility.

Cyber operations during armed conflicts: Implications for international humanitarian law

It is well-established that international law applies to cyber operations. More specifically, IHL and its principles of distinction, proportionality, precaution, military necessity and humanity restrict the use of cyber means and methods during armed conflict. Further discussions may however be needed to clarify how IHL applies and whether it is adequate and sufficient or requires further development, building on existing law.

The meeting helped to clarify which areas of humanitarian concern should be the focus of attention. In brief, based on the detailed knowledge available of cyber operations during peacetime, and the somewhat lesser knowledge of cyber operations in times of armed conflict, the following picture emerges.

Distinction in cyber space

First, cyber attacks are not necessarily indiscriminate. As the report illustrates in more detail, cyber tools can be designed to self-propagate or not. Even if they self-propagate and cause cyber security concerns for all those infected, they can be designed to only cause damage to a specific target. While some self-propagating malware that caused indiscriminate harmful effects has made headlines, many cyber operations have in fact been rather discriminate from a technical perspective (which does not mean they were lawful).

Furthermore, certain types of cyber attacks, such as those that would aim to cause physical damage to industrial control systems, require custom-made cyber tools. In many cases this would also effectively hamper the ability to carry such attacks out in a large-scale, indiscriminate manner.

This is important from an IHL perspective, because contrary to the assumption often heard that the principle of distinction might have become meaningless in cyberspace because of the interconnectivity that characterizes it, not all offensive cyber tools are inherently indiscriminate. On the contrary, they may well be very precisely tailored and create effects on specific targets only.

https://doi.org/10.1017/S181638312000048X Published online by Cambridge University Press

Highlighting the potential human cost

Secondly, and of equal importance, it is nonetheless clear that cyber tools can cause substantial damage and can be – and have sometimes been – indiscriminate, and that certain systems are particularly at risk, first and foremost, perhaps, health-care systems. Moreover, the threats that can be observed have been evolving faster than anticipated, in particular regarding attacks against industrial systems. Finally, much is still unknown in terms of the rapid evolution of the technology, the capabilities and the tools developed by the most sophisticated actors, and the extent to which the increased use of cyber operations during armed conflicts might be different from the trends observed so far. In other words, while the risk of human cost based on current observations does not appear extremely high, especially considering the destruction and suffering that conflicts always cause, the evolution of cyber operations still merits close attention due to existing uncertainties and the rapid pace of change.

Legal protection through IHL

Many of the attacks described in the report targeted or indiscriminately affected civilian infrastructure. In the view of the ICRC, if carried out in times of armed conflict, such attacks would be prohibited. First of all, direct attacks against civilian infrastructure and indiscriminate attacks would be prohibited. Secondly, even if the infrastructure or some parts of it had become military objectives (such as a part of an electricity grid), IHL would require that only this part be attacked, and that there be no excessive damage to the remaining civilian parts. Thirdly, IHL would require parties to the conflict to take all feasible precautions to avoid or at least minimize incidental harm to civilians and civilian objects. Finally, even when they do not amount to attacks under IHL,[3] such operations might be prohibited by the specific protection afforded by IHL to medical facilities or objects indispensable to the survival of the population. These are powerful protections that remain entirely relevant in view of the technical characteristics of cyber operations. For IHL to truly provide legal protection to civilians against the effects of cyber warfare, however, States must commit to its applicability and to an interpretation of its rules that is effective for the protection of civilians and civilian infrastructure. In particular, it would require a clear recognition that cyber operations which impair the functionality of civilian infrastructure are subject to the rules governing attacks under IHL.[4] This report will hopefully help to illustrate the need for such an interpretation in order to ensure that civilian infrastructure is protected.

3 Under IHL, "attack" has a specific meaning which would not encompass all cyber operations that are referred to as cyber attacks in a colloquial sense.

4 See ICRC, above note 1, p. 41.

https://doi.org/10.1017/S181638312000048X Published online by Cambridge University Press

Avenues that could be explored to reduce the potential human cost of cyber operations

Cyber security measures

Beyond the restraints imposed by IHL upon those carrying out cyber operations, it is critical to enhance the cyber security posture and resilience of the actors potentially affected. While cyber security and defence are constantly improving, older systems with outdated or even nonexistent cyber security are particularly vulnerable to cyber attacks and will remain a concern in the years to come. Both the public and private sectors have a role to play through industry standards and legal regulation.

In the health-care sector, for instance, the regulatory environment should be adapted to the increased risk, such as through standardization requirements, with a view to ensuring resilience in the event of a cyber attack. Cyber security needs to be taken into account in the design and development of medical devices and updated throughout their lifetime, no matter how long they last. Similarly, for industrial control systems, industry standards, whether imposed or self-imposed, are critical. This includes reporting incidents and sharing information between trusted partners.

In terms of IHL, parties to armed conflicts must take all feasible precautions to protect civilians and civilian objects under their control against the effects of attack. This is one of the few IHL obligations that States must already implement in peacetime.

Disclosing vulnerabilities

The preferred option for enhancing the safety of cyberspace should be disclosing vulnerabilities to the appropriate software developer so that those vulnerabilities can be fixed. Some States have therefore put in place equity processes to balance competing interests and risks and decide whether to disclose the vulnerabilities they identify.

Measures to prevent proliferation

Those who develop cyber weapons should consider creating obstacles in order to make repurposing difficult and expensive. While it is hardly possible from a technical standpoint to guarantee that malware cannot be repurposed, methods like encrypting its payload and including obstacles in different components of the code, for example, could raise the bar in terms of the expertise required to re-engineer malicious tools. While there is currently no express obligation under IHL to create obstacles to the repurposing of cyber tools, this could prevent at least some actors from doing so and therefore reduce the risk of subsequent misuse that their proliferation entails. The unique way in which cyber tools proliferate also raises the question of whether existing law is adequate or sufficient to address this phenomenon.

https://doi.org/10.1017/S181638312000048X Published online by Cambridge University Press

Marking of certain civilian infrastructure

Another avenue, which builds on existing international law, could be to create a "digital watermark" to identify certain actors or infrastructure in cyberspace that must be protected (such as objects that enjoy specific protection under IHL). The aim would be to help their identification and prevent them from being targeted during armed conflicts. The potentially positive effects in terms of protection against unintended harm by law-abiding actors would however need to be balanced against the risk of disclosing information on critical infrastructure to potential adversaries, including criminals. The prospects of positive effects might depend in part on attribution becoming easier.

Improving attribution and accountability

Finally, enhanced attribution capacities would help ensure that actors who violate international law in cyberspace can be held accountable, which is a means to strengthen compliance with the law and more generally encourage responsible behaviour in cyberspace.

Way forward

The use of cyber operations in armed conflict is likely to continue and might remain shrouded in secrecy. Analyzing its consequences is a complex and long-term endeavour that requires multidisciplinary expertise and interaction with a wide variety of stakeholders.

Building upon the conclusions reached at the expert meeting, the ICRC would like to pursue dialogue with governments, experts and the IT sector. It looks forward to the feedback to this report in order to continue to follow the evolution of cyber operations, in particular during armed conflicts, and their potential human cost, to explore avenues that could reduce them, and to work towards a consensus on the interpretation of existing IHL rules, and potentially the development of complementary rules that afford effective protection to civilians.

Available at: www.icrc.org/en/document/potential-human-cost-cyber-operations.

***Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control*, August 2019**

Executive summary

The ICRC has emphasized the need to maintain human control over weapon systems and the use of force, to ensure compliance with international law and to satisfy ethical concerns. This approach has informed the ICRC's analysis of the legal, ethical, technical and operational questions raised by autonomous weapon systems.

https://doi.org/10.1017/S181638312000048X Published online by Cambridge University Press

In June 2018, the ICRC convened a round-table meeting with independent experts in autonomy, artificial intelligence (AI) and robotics to gain a better understanding of the technical aspects of human control, drawing on experience with civilian autonomous systems. This report combines a summary of the discussions at that meeting with additional research, and highlights the ICRC's main conclusions, which do not necessarily reflect the views of the participants. Experience in the civilian sector yields insights that can inform efforts to ensure meaningful, effective and appropriate human control over weapon systems and the use of force.

Autonomous (robotic) systems operate without human intervention, based on interaction with their environment. These systems raise such questions as "How can one ensure effective human control of their functioning?" and "How can one foresee the consequences of using them?" The greater the complexity of the environment and the task, the greater the need for direct human control and the less one can tolerate autonomy, especially for tasks and in environments that involve risk of death and injury to people or damage to property – in other words, safety-critical tasks.

Humans can exert some control over autonomous systems – or specific functions – through supervisory control, meaning "human-on-the-loop" supervision and the ability to intervene and deactivate. This requires the operator to have:

- situational awareness;
- enough time to intervene;
- a mechanism through which to intervene (a communication link or physical controls) in order to take back control, or to deactivate the system should circumstances require.

However, human-on-the-loop control is not a panacea, because of such human–machine interaction problems as automation bias, lack of operator situational awareness and the moral buffer.

Predictability and reliability are at the heart of discussions about autonomy in weapon systems, since they are essential to achieving compliance with IHL and avoiding adverse consequences for civilians. They are also essential for military command and control.

It is important to distinguish between reliability – a measure of how often a system fails; and predictability – a measure of how the system will perform in a particular circumstance. Reliability is a concern in all types of complex system, whereas predictability is a particular problem with autonomous systems. There is a further distinction between predictability in a narrow sense of knowing the process by which the system functions and carries out a task, and predictability in a broad sense of knowing the outcome that will result.

It is difficult to ensure and verify the predictability and reliability of an autonomous (robotic) system. Both factors depend not only on technical design but also on the nature of the environment, the interaction of the system with that environment, and the complexity of the task. However, setting boundaries or

https://doi.org/10.1017/S181638312000048X Published online by Cambridge University Press

imposing constraints on the operation of an autonomous system – in particular on the task, the environment, the time frame of operation and the scope of operation over an area – can render the consequences of using such a system more predictable.

In a broad sense, all autonomous systems are unpredictable to a degree because they are triggered by their environment. However, developments in the complexity of software control systems – especially those based on AI and machine learning – add unpredictability in the narrow sense that the process by which the system functions is unpredictable.

The "black box" manner in which many machine learning systems function makes it difficult – and in many cases impossible – for the user to know how the system reaches its output. Not only are such algorithms unpredictable but they are also subject to bias, whether by design or in use. Furthermore, they do not provide explanations for their outputs, which seriously complicates establishing trust in their use and exacerbates the already significant challenges of testing and verifying the performance of autonomous systems. And the vulnerability of AI and machine learning systems to adversarial tricking or spoofing amplifies the core problems of predictability and reliability.

Computer vision and image recognition are important applications of machine learning. These applications use deep neural networks (deep learning), of which the functioning is neither predictable nor explainable, and such networks can be subject to bias. More fundamentally, machines do not see like humans. They have no understanding of meaning or context, which means they make mistakes that a human never would.

It is significant that industry standards for civilian safety-critical autonomous robotic systems – such as industrial robots, aircraft autopilot systems and self-driving cars – set stringent requirements regarding human supervision, intervention and deactivation, or fail-safe; predictability and reliability; and operational constraints. Leading developers of AI and machine learning have stressed the need to ensure human control and judgement in sensitive applications – and to address safety and bias – especially where applications can have serious consequences for people's lives.

Civilian experience with autonomous systems reinforces and expands some of the ICRC's viewpoints and concerns regarding autonomy in the critical functions of weapon systems. The consequences of using autonomous weapon systems are unpredictable because of uncertainty for the user regarding the specific target, and the timing and location of any resulting attack. These problems become more pronounced as the environment or the task become more complex, or freedom of action in time and space increases. Human-on-the-loop supervision and intervention and the ability to deactivate are absolute minimum requirements for countering this risk, but the system must be designed to allow for meaningful, timely, human intervention – and even that is no panacea.

All autonomous weapon systems will always display a degree of unpredictability stemming from their interaction with the environment. It might be possible to mitigate this to some extent by imposing operational constraints on the task, the time frame of operation, the scope of operation over an area and the

https://doi.org/10.1017/S181638312000048X Published online by Cambridge University Press

environment. However, the use of software control based on AI – and especially machine learning, including applications in image recognition – brings with it the risk of inherent unpredictability, lack of explainability and bias. This heightens the ICRC's concerns regarding the consequences of using AI and machine learning to control the critical functions of a weapon system and raises questions about their use in decision support systems for targeting.

This review of technical issues highlights the difficulty of exerting human control over autonomous (weapon) systems and shows how AI and machine learning could exacerbate this problem exponentially. Ultimately it confirms the need for States to work urgently to establish limits on autonomy in weapon systems.

Further, this review reinforces the ICRC's view that States should agree on the type and degree of human control required to ensure compliance with international law and to satisfy ethical concerns, while also underlining its doubts that autonomous weapon systems could be used in compliance with IHL in all but the narrowest of scenarios and the simplest of environments.

Available at: www.icrc.org/en/document/autonomy-artificial-intelligence-and-robotics-technical-aspects-human-control.

Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control, June 2020

ICRC and SIPRI, by Vincent Boulanin, Neil Davison, Netta Goussac and Moa Peldán Carlsson

Executive summary[5]

The challenges posed by autonomous weapon systems (AWS) are the focus of an intergovernmental discussion under the framework of the United Nations Convention on Certain Conventional Weapons (CCW). Despite enduring disagreements on whether additional regulation is needed, and in what form, there is emerging consensus among States that autonomy in weapon systems cannot be unlimited: humans must "retain" and "exercise" responsibility for the use of weapon systems and the use of force in armed conflict. This report explores the difficult question of how that principle must be applied in practice. It offers an in-depth discussion on the type and degree of control that humans need to exercise over AWS, in light of legal requirements, ethical concerns and operational considerations. It provides policy-makers with practical guidance on how measures for human control should form the basis of internationally agreed limits on AWS, whether rules, standards or best practices.

The report is the result of a joint project of the ICRC and the Stockholm International Peace Research Institute (SIPRI). Chapter 1 introduces the context and conceptual approach. Chapter 2 explores the legal, ethical and operational

5 This executive summary © SIPRI 2020, reproduced with permission.

https://doi.org/10.1017/S181638312000048X Published online by Cambridge University Press

perspectives on human control. Chapter 3 provides practical guidance on the type, degree and combination of control measures needed for compliance with IHL and to address ethical concerns, while taking into account military operational considerations. Chapter 4 presents the key findings and recommendations for policy-makers.

A core problem with AWS is that they are triggered by the environment, meaning that the user does not know, or choose, the specific target, timing and/or location of the resulting application of force. This process by which AWS function and the associated unpredictability in the consequences of their use can raise serious risks for civilians and challenges for compliance with IHL, as well as fundamental ethical concerns about the role of humans in life-and-death decisions, and challenges for military command and control.

A key question, therefore, is what limits are needed on AWS to address these challenges. An examination of the legal, ethical and operational requirements for human control indicates the need for a combination of three types of control measures:

1. *Controls on the weapon system's parameters of use*, including measures that restrict the type of target and the task the AWS is used for; place temporal and spatial limits on its operation; constrain the effects of the AWS; and allow for deactivation and fail-safe mechanisms.
2. *Controls on the environment*, namely, measures that control or structure the environment in which the AWS is used (e.g. using the AWS only in environments where civilians and civilian objects are not present, or excluding their presence for the duration of the operation).
3. *Controls through human–machine interaction*, such as measures that allow the user to supervise the AWS and to intervene in its operation where necessary.

These control measures can help to reduce or at least compensate for the unpredictability inherent in the use of AWS and to mitigate the risks involved, in particular for civilians. From a legal perspective, a user must exercise sufficient control to have reasonable certainty about the effects of an AWS when used in an attack and to be able to limit them as required by IHL. Ethical considerations may demand additional constraints, especially given concerns with AWS designed or used against persons.

The report concludes with five recommendations. First, States should focus their work on determining how measures needed for human control apply in practice. Since these three types of control measures are not tied to specific technologies, they provide a robust normative basis applicable to the regulation of both current and future AWS.

Second, measures for human control should inform any development of internationally agreed limits on AWS, whether new rules, standards or best practices. This work must be guided by the legal, ethical and operational requirements for human control. Any normative development should also focus on human obligations and responsibilities, not on technological fixes, so as to remain relevant and practical, and adaptable to future technological developments.

https://doi.org/10.1017/S181638312000048X Published online by Cambridge University Press

Third, States should clarify where IHL rules already set constraints on the development and use of AWS, and where new rules, standards and best practice guidance may be needed.

Fourth, any new rules, standards and best practices must build on existing limits on autonomy under IHL, and should draw on existing practice. It is likely that new rules, standards and best practice guidance can be most effectively articulated in terms of limits on specific types of AWS and on the manner and circumstances of their use, and requirements for human supervision and intervention.

Fifth, human control criteria should be considered in the study, research and development, and acquisition of new weapon systems.

Available at: www.icrc.org/en/document/limits-autonomous-weapons.

https://doi.org/10.1017/S181638312000048X Published online by Cambridge University Press