ICRC

Ethics and autonomous weapon systems: An ethical basis for human control?

International Committee of the Red Cross (ICRC)

Geneva, 3 April 2018

EXECUTIVE SUMMARY

In the view of the International Committee of the Red Cross (ICRC), **human control must be maintained** over weapon systems and the use of force to ensure compliance with international law and to satisfy ethical concerns, and **States must work urgently to establish limits on autonomy in weapon systems.**

In August 2017, the ICRC convened a round-table meeting with independent experts to explore the ethical issues raised by autonomous weapon systems and the **ethical dimension of the requirement for human control**. This report summarizes discussions and highlights the ICRC's main conclusions.

The **fundamental ethical question** is whether the principles of humanity and the dictates of the public conscience can allow human decision-making on the use of force to be effectively substituted with computer-controlled processes, and **life-and-death decisions to be ceded to machines.**

It is clear that ethical decisions by States, and by society at large, have preceded and motivated the development of new international legal constraints in warfare, including constraints on weapons that cause unacceptable harm. In international humanitarian law, notions of humanity and public conscience are drawn from the **Martens Clause.** As a potential marker of the public conscience, opinion polls to date suggest a general opposition to autonomous weapon systems – with autonomy eliciting a stronger response than remote-controlled systems.

Ethical issues are at the heart of the debate about the acceptability of autonomous weapon systems. It is precisely anxiety about the loss of human control over weapon systems and the use of force that goes **beyond questions of the compatibility of autonomous weapon systems with our *laws*** to encompass **fundamental questions of acceptability to our *values***. A prominent aspect of the ethical debate has been a focus on autonomous weapon systems that are designed to kill or injure humans, rather than those that destroy or damage objects, which are already employed to a limited extent.

The **primary ethical argument *for* autonomous weapon systems has been results-oriented**: that their potential precision and reliability might enable better respect for both international law and human ethical values, resulting in fewer adverse humanitarian consequences. As with other weapons, such characteristics would depend on both the design-dependent effects and the way the weapons were used. A secondary argument is that they would help fulfil the duty of militaries to protect their own forces – a quality not unique to autonomous weapon systems.

While there are concerns regarding the technical capacity of autonomous weapons systems to function within legal and ethical constraints, the **enduring ethical arguments *against* these weapons are those that transcend context** – whether during armed conflict or in peacetime – and transcend technology – whether simple or sophisticated.

The importance of **retaining human agency – and intent – in decisions to use force,** is one of the central ethical arguments for limits on autonomy in weapon systems. Many take the view that decisions to kill, injure and destroy must not be delegated to machines, and that humans must be

present in this decision-making process sufficiently to preserve a direct link between the intention of the human and the eventual operation of the weapon system.

Closely linked are **concerns about a loss of human dignity**. In other words, it matters not just *if* a person is killed or injured but *how* they are killed or injured, including the process by which these decisions are made. It is argued that, if human agency is lacking to the extent that machines have effectively, and functionally, been delegated these decisions, then it undermines the human dignity of those combatants targeted, and of civilians that are put at risk as a consequence of legitimate attacks on military targets.

The need for human agency is also linked to **moral responsibility and accountability** for decisions to use force. These are human responsibilities (both ethical and legal), which **cannot be transferred to inanimate machines, or computer algorithms**.

Predictability and **reliability** in using an autonomous weapon system are ways of connecting human agency and intent to the eventual consequences of an attack. However, as weapons that self-initiate attacks, **autonomous weapon systems all raise questions about predictability**, owing to varying degrees of uncertainty as to exactly when, where and/or why a resulting attack will take place. The application of **AI and machine learning to targeting functions raises fundamental questions of inherent unpredictability.**

Context also affects ethical assessments. Constraints on the **timeframe of operation** and **scope of movement over an area** are key factors, as are the **task** for which the weapon is used and the **operating environment**. However, perhaps the most important factor is the **type of target,** since core ethical concerns about human agency, human dignity and moral responsibility are **most acute in relation to the notion of anti-personnel autonomous weapon systems that target humans directly**.

From the ICRC's perspective**, ethical considerations parallel the requirement for a minimum level of human control** over weapon systems and the use of force to ensure legal compliance. From an ethical viewpoint, "**meaningful", "effective" or "appropriate" human control would be the type and degree of control that preserves human agency and upholds moral responsibility in decisions to use force**. This requires a sufficiently direct and close connection to be maintained between the human intent of the user and the eventual consequences of the operation of the weapon system in a specific attack.

Ethical and legal considerations may demand some similar constraints on autonomy in weapon systems, so that meaningful human control is maintained – in particular, with respect to: **human supervision and the ability to intervene and deactivate**; technical requirements for **predictability** and **reliability** (including in the algorithms used); and **operational constraints** on the task for which the weapon is used, the type of target, the operating environment, the timeframe of operation and the scope of movement over an area.

However, the **combined and interconnected ethical concerns** about loss of human agency in decisions to use force, diffusion of moral responsibility and loss of human dignity **could have the most far-reaching consequences, perhaps precluding the development and use of anti-personnel autonomous weapon systems**, and even limiting the applications of anti-materiel systems, depending on the risks that destroying materiel targets present for human life.

CONTENTS

1. INTRODUCTION

Since 2011, the ICRC has been engaged in debates about autonomous weapon systems, holding international expert meetings with States and independent experts in March 2014[1] and March 2016,[2] and contributing to discussions at the United Nations Convention on Certain Conventional Weapons (CCW) since 2014.

The **ICRC's position is that States must establish limits on autonomy in weapon systems** to ensure compliance with international humanitarian law and other applicable international law, and to satisfy ethical concerns. It has called on States to determine where these limits should be placed by assessing the **type and degree of human control** required in the use of autonomous weapon systems (broadly defined as weapons with autonomy in their critical functions of selecting and attacking targets)[3] for legal compliance and ethical acceptability.[4]

As part of continuing reflections, the **ICRC convened a two-day round-table meeting** with independent experts to consider the ethical issues raised by autonomous weapon systems and the **ethical dimension of the requirement for human control** over weapon systems and the use of force.[5] This report summarizes discussions at the meeting, supplemented by additional research. The report highlights key themes and conclusions from the perspective of the ICRC, and these do not necessarily reflect the views of the participants.

For the ICRC, the **fundamental question at the heart of ethical discussions** is whether, irrespective of compliance with international law, the principles of humanity and the dictates of the public conscience can allow human decision-making on the use of force to be effectively substituted with computer-

[1] ICRC, *Autonomous weapon systems: Technical, military, legal and humanitarian aspects*, 2014 – report of an expert meeting: https://www.icrc.org/en/document/report-icrc-meeting-autonomous-weapon-systems-26-28-march-2014.

[2] ICRC, *Autonomous weapon systems: Implications of increasing autonomy in the critical functions of weapons*, 2016 – report of an expert meeting: https://www.icrc.org/en/publication/4283-autonomous-weapons-systems.

[3] The ICRC's working definition of an autonomous weapon system is: "*Any weapon system with autonomy in its critical functions. That is, a weapon system that can select (i.e. search for or detect, identify, track, select) and attack (i.e. use force against, neutralize, damage or destroy) targets without human intervention*." This definition encompasses a limited number of existing weapons, such as: anti-materiel weapon systems used to protect ships, vehicles, buildings or areas from incoming attacks with missiles, rockets, artillery, mortars or other projectiles; and some loitering munitions. There have been reports that some anti-personnel "sentry" weapon systems have autonomous modes. However, as far as is known to the ICRC, "sentry" weapon systems that have been deployed still require human remote authorization to launch an attack (even though they may identify targets autonomously). See: ICRC, *Autonomous weapon systems: Implications of increasing autonomy in the critical functions of weapons*, *op. cit.* (footnote 2), 2016, pp. 11–12.

[4] ICRC, *Statement to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on "Lethal Autonomous Weapon Systems"*, 15 November 2017: https://www.icrc.org/en/document/expert-meeting-lethal-autonomous-weapons-systems; N Davison, "Autonomous weapon systems under international humanitarian law", in *Perspectives on Lethal Autonomous Weapon Systems, United Nations Office for Disarmament* Affairs *(UNODA) Occasional Papers No. 30*, November 2017: https://www.un.org/disarmament/publications/occasionalpapers/unoda-occasional-papers-no-30-november-2017; ICRC, *Views of the ICRC on autonomous weapon systems*, 11 April 2016: https://www.icrc.org/en/document/views-icrc-autonomous-weapon-system.

[5] The event was entitled "*Ethics and autonomous weapon systems: An ethical basis for human control?*" and was held at the Humanitarium, International Committee of the Red Cross (ICRC), Geneva, on 28 and 29 August 2017. With thanks to the following experts for their participation: Joanna Bryson (University of Bath, UK); Raja Chatila (Institut des Systèmes Intelligents et de Robotique, France); Markus Kneer (University of Zurich, Switzerland); Alexander Leveringhaus (University of Oxford, UK); Hine-Wai Loose (United Nations Office for Disarmament Affairs, Geneva); AJung Moon (Open Roboethics Institute, Canada); Bantan Nugroho (United Nations Office for Disarmament Affairs, Geneva); Heather Roff (Arizona State University, USA); Anders Sandberg (University of Oxford, UK); Robert Sparrow (Monash University, Australia); Ilse Verdiesen (Delft University of Technology, Netherlands); Kerstin Vignard (United Nations Institute for Disarmament Research); Wendell Wallach (Yale University, US); and Mary Wareham (Human Rights Watch). The ICRC was represented by: Kathleen Lawand, Neil Davison and Anna Chiapello (Arms Unit, Legal Division); Fiona Terry (Centre for Operational Research and Experience); and Sasha Radin (Law and Policy Forum). Report prepared by Neil Davison, ICRC.

controlled processes, and **life-and-death decisions to be ceded to machines.** The ICRC's concerns reflect the sense of deep discomfort over the idea of any weapon system that places the use of force beyond human control.[6] And yet, important questions remain: at what point have decisions effectively, or functionally, been delegated to machines? What type and degree of human control are required, and in which circumstances, to satisfy ethical concerns? These are questions with profound implications for the future of warfare and humanity, and all States, as well as the military, scientists, industry, civil society and the public, have a stake in determining the answers.

2. THE PRINCIPLES OF HUMANITY AND THE DICTATES OF THE PUBLIC CONSCIENCE

2.1 Ethics and the law

Ethics and law are intimately linked, especially where the purpose of the law – such as international humanitarian law and international human rights law – is to protect persons. This relationship can provide insights into how considerations of humanity and public conscience drive legal development.

The regulation of any conduct of hostilities, including regulating the choice of weapons, starts with a societal decision of what is acceptable or unacceptable behaviour, what is right and wrong. Subsequent **legal restrictions are**, therefore, **a social construct, shaped by societal and ethical perceptions**. These determinations evolve over time; what was considered acceptable at one point in history is not necessarily the case today.[7] However, some codes of behaviour in warfare have endured for centuries – for example, the unacceptability of killing women and children, and of poisoning.

It is clear that ethical decisions by States, and by society at large, have preceded and motivated the development of new international legal constraints in warfare, and that in the face of new developments not specifically foreseen or not clearly addressed by existing law, **contemporary ethical concerns can go beyond what is already codified in the law**. This highlights the importance of not reducing debates about autonomous weapon systems, or other new technologies of warfare, solely to legal compliance.

2.2 The Martens Clause

In international humanitarian law, **notions of humanity and public conscience are drawn from the Martens Clause**, a provision that first appeared in the Hague Conventions of 1899 and 1907, was later incorporated in the 1977 Additional Protocols to the Geneva Conventions, and is considered customary law. It provides that, in cases not covered by existing treaties, civilians and combatants remain under the protection and authority of the principles of humanity and the dictates of the public conscience.[8] The Martens Clause prevents the assumption that anything that is not explicitly prohibited by relevant

[6] ICRC, *Statement to the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on "Lethal Autonomous Weapons Systems"*, 13 April 2015: https://www.icrc.org/en/document/lethal-autonomous-weapons-systems-LAWS.
[7] For example, among conventional weapons: expanding bullets, anti-personnel mines and cluster munitions.
[8] It appears in the preamble to Additional Protocol II and in Article 1(2) of Additional Protocol I: "In cases not covered by this Protocol or by any other international agreements, civilians and combatants remain under the protection and authority of the principles of international law derived from established custom, from the principles of humanity and from dictates of public conscience."

treaties is therefore permitted – **it is a safety net for humanity**. The provision is recognized as being particularly relevant to assessing new technologies and new means and methods of warfare.[9]

There is debate over whether the Martens Clause constitutes a legally-binding yardstick against which the lawfulness of a weapon must be measured, or rather an ethical guideline. Nevertheless, it is clear that considerations of humanity and public conscience have driven the evolution of international law on weapons, and these notions have triggered the negotiation of specific treaties to prohibit or limit certain weapons, as well as underlying the development and implementation of the rules of international humanitarian law more broadly.[10]

2.3 The public conscience in practice

In the development of international humanitarian law on weapons there is a strong ethical narrative to be found in the words used by States, the ICRC (mandated to uphold international humanitarian law) and civil society in raising concerns about **weapons that cause, or have the potential to cause, unacceptable harm**. For example, regarding weapons that cause **superfluous injury or unnecessary suffering for combatants**, in 1918, the ICRC, in calling for a prohibition of chemical weapons, described them as "barbaric weapons", an "appalling method of waging war", and appealed to States' "feeling of humanity".[11] In advocating for a prohibition of blinding laser weapons, the ICRC appealed to the "conscience of humanity" and later welcomed the 1995 Protocol IV to the Convention on Certain Conventional Weapons (CCW) as a "victory of civilization over barbarity".[12]

Likewise, addressing **weapons that strike blindly, indiscriminately affecting civilians**, the ICRC expressed an ethical revulsion over the "landmine carnage" and "appalling humanitarian consequences" of anti-personnel mines in debates leading to the prohibition of these weapons in 1997.[13] The recent Treaty on the Prohibition of Nuclear Weapons, adopted in July 2017 by a group of 122 States, recognizes that the use of nuclear weapons would be "abhorrent to the principles of humanity and the dictates of public conscience".[14] The ethical underpinnings of restrictions in international humanitarian law on the use of certain weapons are not in dispute.

Civil society, medical, scientific and military experts, and the ICRC and other components of the International Red Cross and Red Crescent Movement, have played a key role in raising the attention of States to the unacceptable harm caused by certain weapons, such as anti-personnel mines and cluster munitions, building on evidence collected by those treating victims. Engagement in these endeavours by military veterans and religious figures, appeals to political leaders and parliamentarians, the testimony of victims and communication of concerns to the public were central to securing these prohibitions. In some debates, such as on blinding laser weapons, reflections by the

---

[9] International Court of Justice, *Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion*, ICJ Reports, 1996, para. 78.

[10] K Lawand and I Robinson, "Development of treaties limiting or prohibiting the use of certain weapons: the role of the International Committee of the Red Cross", in R Geiss, A Zimmermann and S Haumer (eds.), *Humanizing the laws of war: the Red Cross and the development of international humanitarian law*, Cambridge University Press, 2017, pp. 141–184; M Veuthey, "Public Conscience in International Humanitarian Law", in D Fleck (ed.), *Crisis Management and Humanitarian Protection*, Berliner Wissenschafts-Verlag, Berlin, 2004, pp. 611–642.

[11] ICRC, *World War I: the ICRC's appeal against the use of poisonous gases*, 1918: https://www.icrc.org/eng/resources/documents/statement/57jnqh.htm.

[12] L Doswald-Beck, "New Protocol on Blinding Laser Weapons", *International Review of the Red Cross*, No. 312, 1996: https://www.icrc.org/eng/resources/documents/article/other/57jn4y.htm.

[13] P Herby and K Lawand, "Unacceptable Behaviour: How Norms are Established", in J Williams, S Goose and M Wareham (eds.), *Banning Landmines: Disarmament, Citizen Diplomacy and Human Security,* Lanham, MD: Rowman & Littlefield Publishers, 2008, p. 202.

[14] UN General Assembly, Treaty on the Prohibition of Nuclear Weapons, preamble, A/CONF.229/2017/8, 7 July 2017.

military on the risks for their own soldiers were critical. All these various activities can be seen, in some way, as a demonstration of the public conscience.[15]

3. THE ETHICAL DEBATE ON AUTONOMOUS WEAPON SYSTEMS

Ethical questions about autonomous weapon systems have sometimes been viewed as secondary concerns. Many States have tended to be more comfortable discussing whether new weapons can be developed and used in compliance with international law, particularly international humanitarian law, and with the assumption that the primary factors that limit the development and use of autonomous weapon systems are legal and technical.

However, for many experts and observers, and for some States, **ethics** – the "moral principles that govern a person's behaviour or the conducting of an activity"[16] – **are at the heart of what autonomous weapon systems mean for the human conduct of warfare, and the use of force more broadly**. It is precisely anxiety about the loss of human control over this conduct that goes beyond questions of the compatibility of autonomous weapon systems with our *laws* to encompass fundamental questions of acceptability to our *values*.

Ethical concerns over delegating life-and-death decisions, and reflections on the importance of the Martens Clause, have been raised in different quarters, including by: more than 30 States during CCW meetings,[17] a UN Special Rapporteur at the Human Rights Council,[18] Human Rights Watch[19] (and the Campaign to Stop Killer Robots), the ICRC,[20] the United Nations Institute for Disarmament Research (UNIDIR),[21] academics and think-tanks, and, increasingly, among the scientific and technical communities.[22]

Discussions on autonomous weapon systems have generally **acknowledged the necessity for some degree of human control over weapons and the use for force**, whether for legal, ethical or military operational reasons (States have not always made clear for which reasons, or combination thereof).[23] It is clear, however, that the points at which human control is located in the development and

---

[15] K Lawand and I Robinson, *op. cit*. (footnote 10), 2017.

[16] *Oxford Dictionary of English*: https://en.oxforddictionaries.com/definition/ethics.

[17] Including: Algeria, Argentina, Austria, Belarus, Brazil, Cambodia, Costa Rica, Cuba, Ecuador, Egypt, France, Germany, Ghana, Holy See, India, Kazakhstan, Mexico, Morocco, Nicaragua, Norway, Pakistan, Panama, Peru, Republic of Korea, Sierra Leone, South Africa, Sri Lanka, Sweden, Switzerland, Turkey, Venezuela, Zambia and Zimbabwe.

[18] Human Rights Council, *Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, Christof Heyns*, A/HRC/23/47, 9 April 2013.

[19] Human Rights Watch, *Losing Humanity: The Case against Killer Robots*, 19 November 2012.

[20] ICRC, Statement to CCW Meeting of Experts on "Lethal Autonomous Weapons Systems", 13–17 April 2015: https://www.icrc.org/en/document/lethal-autonomous-weapons-systems-LAWS.

[21] UNIDIR, *The Weaponization of Increasingly Autonomous Technologies: Considering Ethics and Social Values*, 2015.

[22] Future of Life Institute, *Autonomous Weapons: an Open Letter from AI & Robotics Researchers*, 28 July 2015; Future of Life Institute, *An Open Letter to the United Nations Convention on Certain Conventional Weapons*, 21 August 2017.

[23] United Nations, *Report of the 2017 Group of Governmental Experts on "Lethal Autonomous Weapons Systems" (LAWS)*, CCW/GGE.1/2017/CRP.1, 20 November 2017, p.7: "The importance of considering LAWS ["Lethal Autonomous Weapon Systems"] in relation to human involvement and the human-machine interface was underlined. The notions that human control over lethal targeting functions must be preserved, and that machines could not replace humans in making decisions and judgements, were promoted. Various related concepts, including, *inter alia*, meaningful and effective human control, appropriate human judgement, human involvement and human supervision, were discussed."
United Nations, *Recommendations to the 2016 Review Conference Submitted by the Chairperson of the Informal Meeting of Experts*, November 2016, p. 1: "[V]iews on appropriate human involvement with regard to lethal force and the issue of delegation of its use are of critical importance to the further consideration of LAWS amongst the High Contracting Parties and should be the subject of further consideration".

deployment, and exercised in the use, of a weapon with autonomy in the critical functions of selecting and attacking targets may be central to determining whether this control is "meaningful", "effective" or "appropriate" from an ethical perspective (and a legal one).

A prominent aspect of the ethical debate has been a **focus on "lethal autonomy" or "killer robots**" – **implying weapon systems that are designed to kill or injure humans**, rather than autonomous weapon systems that destroy or damage objects, which are already employed to a limited extent.[24] This is despite the fact that some anti-materiel weapons can also result in the death of humans either directly (humans inside objects, such as buildings, vehicles, ships and aircraft) or indirectly (humans in proximity to objects), and that even the use of non-kinetic weapons – such as cyber weapons – can result in kinetic effects and in human casualties. Of course, **autonomy in the critical functions of selecting and attacking targets is a feature that could, in theory, be applied to any weapon system**.

Ethical discussions have also **transcended the context-dependent legal bounds of international humanitarian law and international human rights law**. Ethical concerns, relevant in all circumstances, have been at the centre of warnings by UN Special Rapporteur Christof Heyns that "allowing LARs [Lethal Autonomous Robots] to kill people may denigrate the value of life itself",[25] and by Human Rights Watch that "fully autonomous weapons" would "cross a moral threshold" because of "the lack of human qualities necessary to make a moral decision, the threat to human dignity and the absence of moral agency".[26]

3.1 Main ethical arguments

Nevertheless, ethical arguments have been made both *for* and *against* autonomous weapon systems, reflecting, to a certain extent, the different emphases of consequentialist (results-focused) and deontological (process-focused) approaches. The **primary argument *for* these weapons has been an assertion that they might enable better respect for both international law and human ethical values** by enabling greater precision and reliability than weapon systems controlled directly by humans, and therefore would result in less adverse humanitarian consequences for civilians.[27] This type of argument has been made in the past for other weapon systems, including, most recently, for armed drones, and it is important to recognize that such characteristics are not inherent to a weapon system but depend on both the design-dependent effects and the way the weapon system is used.[28]

Another ethical argument that has been made *for* autonomous weapon systems is that **they help fulfil the duty of militaries to protect their soldiers** by removing them from harm's way. However, since this can equally apply to remote-controlled and remotely-delivered weapons, it is not a convincing argument for autonomy in targeting *per se*, apart from, perhaps, in scenarios where human soldiers cannot respond quickly enough to an incoming threat, such as in missile and close-in air defence.

---

[24] See footnote 3 on existing autonomous weapon systems. Although the use of anti-materiel systems has not been without its problems and accidents – see, for example: J Hawley, *Automation and the Patriot Air and Missile Defense System*, Center for a New American Security (CNAS), 25 January 2017.

[25] Human Rights Council, *op. cit.* (footnote 18), 2013, p. 20.

[26] Human Rights Watch, *Making the Case: The Dangers of Killer Robots and the Need for a Pre-emptive Ban*, 9 December 2016.

[27] See, for example on ethical compliance: R Arkin "Lethal Autonomous Systems and the Plight of the Non-combatant", in *AISIB Quarterly*, July 2013. And on legal compliance: United States, *Autonomy in Weapon Systems*, Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on "Lethal Autonomous Weapon *Systems"*, CCW/GGE.1/2017/WP.6, 10 November 2017, pp. 3–4.

[28] For example, remote-controlled armed drones with precision-guided munitions may offer the potential for greater precision and therefore less risk of indiscriminate effects. However, if the information about the target is inaccurate, targeting practices are too generalized, or protected persons or objects are deliberately, or accidentally, attacked, then the potential for precision offers no protection in itself.

Ethical arguments *against* autonomous weapon systems can generally be divided into two forms: objections based on the limits of technology to function within legal constraints and ethical norms;[29] and ethical objections that are independent of technological capability.[30]

Given that technology trajectories are hard to predict, it is the second category of ethical arguments that may be the most interesting for current policy debates. Do autonomous weapon systems raise any universal ethical concerns? Among the main issues in this respect are:

- **removing human agency from decisions to kill, injure and destroy**[31] – decisions to use force – leading to a **responsibility gap** where humans cannot uphold their moral responsibility[32]
- **undermining the human dignity** of those combatants who are targeted,[33] and of civilians who are put at risk of death and injury as a consequence of attacks on legitimate military targets
- **further increasing human distancing** – physically and psychologically – from the battlefield, enhancing existing asymmetries and making the use of violence easier or less controlled.[34]

3.2 Human agency in decisions to use force

In ethical debates, there seems to be wide acknowledgement of the importance of **retaining human agency**[35] **– and associated intent – in decisions to use force, particularly in decisions to kill, injure and destroy**. In other words, many take the view that "machines must not make life-and-death decisions" and "machines cannot be delegated responsibility for these decisions".[36]

Machines and computer programs, as inanimate objects, do not think, see and perceive like humans. Therefore, some argue, it is difficult to see how human values can be respected if the "decision" to attack a specific target is functionally delegated to a machine. However, there are differing perspectives on the underlying question: at which point have decisions to use force effectively been delegated to a machine? Or, from another perspective: **what limits on autonomy are required to retain sufficient human agency and intent in these decisions**?

---

[29] See, for example: N Sharkey, "The evitability of autonomous robot warfare", *International Review of the Red Cross*, No. 886, 2012.

[30] See, for example: P Asaro, "On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making", *International Review of the Red Cross*, No. 886, 2012;
R Sparrow, "Robots and respect: Assessing the case against Autonomous Weapon Systems", *Ethics and International Affairs*, 30(1), 2016, pp. 93–116; A Leveringhaus, *Ethics and Autonomous Weapon Systems*, Palgrave Macmillan, UK, 2016.

[31] A Leveringhaus, *Ethics and Autonomous Weapon Systems*, *op. cit.* (footnote 30), 2016.

[32] See, for example: R Sparrow, "Killer robots", *Journal of Applied Philosophy*, 24(1), 2007, pp. 62–77; H Roff, "Killing in War: Responsibility, Liability and Lethal Autonomous Robots", in F Allhoff, N Evans and A Henschke (eds.), *Routledge Handbook of Ethics and War: Just War Theory in the 21st Century*, Routledge, UK, 2014.

[33] See, for example: R Sparrow, *op. cit*. (footnote 30), 2016; C Heyns, "Autonomous weapons in armed conflict and the right to a dignified life: An African perspective", *South African Journal on Human Rights*, Vol. 33, Issue 1, 2017, pp. 46–71.

[34] A Leveringhaus, "Distance, weapons technology and humanity in armed conflict", *ICRC Humanitarian Law & Policy Blog*, 6 October 2017: http://blogs.icrc.org/law-and-policy/2017/10/06/distance-weapons-technology-and-humanity-in-armed-conflict.

[35] N Castree, R Kitchin and A Rogers, *A Dictionary of Human Geography*, Oxford University Press, Oxford, 2013: "The capacity possessed by people to act of their own volition."

[36] See footnote 17 listing States that have raised core ethical concerns. For example: "Germany will certainly adhere to the principle that it is not acceptable, that the decision to use force, in particular the decision over life and death, is taken solely by an autonomous system without any possibility for a human intervention." *Statement to CCW Meeting of Experts on "Lethal Autonomous Weapon Systems"*, 11–15 April 2016.

There is a parallel in this debate with landmines, which have been described as "rudimentary autonomous weapon systems".[37] When humans lay landmines they effectively remove themselves from the decision about subsequent attacks on specific people or vehicles. They may know where the landmines are placed but they do not know who, or what, will trigger them, or when they will be triggered. This could be seen as a primitive form of delegating the decision to kill and injure to a machine.

Some argue it is difficult to establish a clear point at which this shift in functional decision-making from human to machine happens, and human agency and intention have been eroded or lost. Rather, it may be more useful, some propose, to agree on the general principle that a minimum level of human control is required in order to retain human agency in these decisions, and then **consider the way in which humans must inject themselves into the decision-making process and at what points, to ensure this control is sufficient** – for example, through human supervision and the ability to intervene and deactivate; technical requirements for predictability and reliability; and operational constraints on the task the weapon is used for, the type of target, the operating environment, the timeframe of operation and the scope of movement over an area.[38]

3.3 Human dignity: process *and* results

Closely linked to the issue of human agency, and concerns about the delegation of decisions to use force, is **human dignity**. The **central argument here is that it matters not just *if* a person is killed and injured but *how* they are killed and injured**. Where a line has been crossed, and machines are effectively making life-and-death "decisions", the argument is that this undermines the human dignity of those targeted, even if they are lawful targets (for example, under international humanitarian law). As Christof Heyns, then UN Special Rapporteur on extrajudicial, summary or arbitrary executions, put it: "to allow machines to determine when and where to use force against humans is to reduce those humans to objects; they are treated as mere targets. They become zeros and ones in the digital scopes of weapons which are programmed in advance to release force without the ability to consider whether there is no other way out, without a sufficient level of deliberate human choice about the matter." [39]

Unlike previous discussions about constraints on weapons (*see Section 2.3*), which have focused on their effects (whether evidence of unacceptable harm or foreseeable effects), the additional ethical concerns with autonomous weapon systems are about *process* as well as *results*. What does this method of using force reveal about the underlying attitude to human life, to human dignity? And, in that sense, these **concerns are particularly relevant to the relationship between combatants in armed conflict, although they are also relevant to civilians**, who must not be targeted, but are, nevertheless, exposed to collateral risks of death and injury from attacks on legitimate military targets.

---

[37] United States Department of Defense, *Department of Defense Law of War Manual*, Section 6.5.9.1, Description and Examples of the Use of Autonomy in Weapon Systems, 2015, p. 328: "Some weapons may have autonomous functions. For example, mines may be regarded as rudimentary autonomous weapons because they are designed to explode by the presence, proximity, or contact of a person or vehicle, rather than by the decision of the operator."
There are different views on whether the complexity of the function delegated to a machine affects this ethical assessment. Some distinguish between an "automated function" (activation, or not, of a landmine) and an "autonomous function" with "choice" (e.g. selecting between different targets), but there are no clear lines between automated and autonomous from a technical perspective, and both can enable functional delegation of decisions. See, for example: ICRC, *Autonomous weapon systems: Implications of increasing autonomy in the critical functions of weapons*, *op. cit.* (footnote 2), 2016, p. 8.

[38] ICRC, *Statement to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on "Lethal Autonomous Weapon Systems"*, *op. cit.* (footnote 4), 15 November 2017.

[39] C Heyns, *Autonomous Weapon Systems: Human rights and ethical issues*, presentation to the CCW Meeting of Experts on "Lethal Autonomous Weapon Systems", 14 April 2016.

For some, autonomous weapon systems conjure up visions of machines being used to kill humans like vermin, and a reduced respect for human life due to a lack of human agency and intention in the specific acts of using force. In this argument, delegating the execution of a *task* to a machine may be acceptable, but delegating the *decision* to kill or injure is not, which means applying human intent to each decision.

There are **strong parallels with the broader societal discussion about algorithmic, and especially artificial intelligence (AI)-driven, decision-making**, including military decision-making[40] (*see also Section 5.1*). Through handing over too much of the functional decision-making process to sensors and algorithms, is there a point at which humans are so far removed in time and space from the acts of selecting and attacking targets that human decision-making is effectively substituted by computer-controlled processes? The concern is that, **if the connection between the human decision to use force and the eventual consequences is too diffuse, then human agency in that decision is weakened and human dignity eroded.**

The counter-argument to an emphasis on process is found in the primary argument *for* autonomous weapons systems (*see Section 3.1*) that they will offer better *results*, posing less risk to civilians by enabling the users to exercise greater precision and discrimination than with human-operated systems. However, claims about reduced risks to civilians – which remain contentious in the absence of supporting evidence – are very much context-specific, whereas ethical questions about loss of human dignity present more of a universal concern, independent of context.

4. RESPONSIBILITY, ACCOUNTABILITY AND TRANSPARENCY

Responsibility and accountability for decisions to use force cannot be transferred to a machine or a computer program.[41] These are human responsibilities – both legal and ethical – which require human agency in the decision-making process (*see Section 3*). Therefore, a closely related ethical concern raised by autonomous weapon systems is the risk of erosion – or diffusion – of responsibility and accountability for these decisions.

One way to address this concern is to assign responsibility to the operator or commander who authorizes the activation of the autonomous weapon system (or programmers and manufacturers, in case of malfunction). This addresses the issue of legal responsibility to some extent, simply by applying a process for holding an individual accountable for the consequences of their actions.[42] And this is how militaries typically address responsibility for operations using existing weapon systems, including, presumably, those with autonomy in their critical functions.

4.1 Implications of autonomy for moral responsibility

For the ethical debate, however, **responsibility is not only a legal concept but also a moral one**. Some argue that, in order for the commander or operator to uphold their moral responsibility in a decision to activate an autonomous weapon system, their **intent needs to be directly linked to the eventual outcome of the resulting attack**. This requires an understanding of how the weapon will function and

[40] D Lewis, G Blum and N Modirzadeh, *War-Algorithm Accountability*, Harvard Law School Program on International Law and Armed Conflict (HLS PILAC), Harvard University, 31 August 2016: https://pilac.law.harvard.edu/waa.

[41] ICRC, *Statement to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on "Lethal Autonomous Weapon Systems"*, *op. cit.* (footnote 4), 15 November 2017.

[42] Although there are still questions around whether a person can be criminally accountable in situations where they lack the required knowledge or intent of how the system will operate once activated, or where there is insufficient evidence to discharge the burden of proof.

the specific consequences of activating it in those circumstances, which is complicated by the uncertainty introduced by autonomy in targeting. Uncertainty brings a risk that the consequences of activating the weapon will not be those intended – or foreseen – by the operator (*see Section 5.2*), which raises both ethical and legal concerns.

An autonomous weapon system – since it selects and attacks targets independently (after launch or activation) – **creates varying degrees of uncertainty as to exactly when, where and/or why the resulting attack will take place**. The key difference between a human or remote-controlled weapon and an autonomous weapon system is that the former involves a human choosing a specific target – or group of targets – to be attacked, connecting their moral (and legal) responsibility to the specific consequences of their actions. In contrast, an **autonomous weapon system self-initiates an attack: it is given a technical description, or a "signature", of a target, and a spatial and temporal area of autonomous operation**. This description might be general ("an armoured vehicle") or even quite specific ("a certain type of armoured vehicle"), but the key issue is that the commander or operator activating the weapon is not giving instructions on a specific target to be attacked ("specific armoured vehicle") at a specific place ("at the corner of that street") and at a specific point in time ("now"). Rather, when activating the autonomous weapon system, by definition, the user will not know exactly which target will be attacked ("armoured vehicles fitting this technical signature"), in which place (within x square kilometres) or at which point in time (during the next x minutes/hours). Thus, it can be argued, this more generalized nature of the targeting decision means the user is not applying their intent to each specific attack.

The potential technical description, or signature, for an enemy combatant is both extremely broad and highly specific (e.g. combatant, fighter or civilian that is directly participating in hostilities but not one that is *hors de combat* or surrendering) and can vary enormously from one moment to the next. It is therefore highly doubtful that a weapon system could be programmed functionally to identify "enemy combatants".[43] But, assuming this might be possible for the sake of argument, if an anti-personnel autonomous weapon system encountered the signature of an enemy combatant it would attack when the signature matches its programming. **A human decision-maker controlling a weapon system in the same circumstances still has a choice**. S/he may decide to attack, or s/he may decide *not* to attack – even if the technical signature fits – including owing to wider ethical considerations in the specific circumstances, which may go beyond whether the combatant is a lawful target.[44] (From a legal perspective, it is important to note that the principles of military necessity and humanity already require that the kind and degree of force used against lawful targets must not exceed what is necessary to accomplish a legitimate military purpose in the circumstances.)[45]

In sum, from an ethical perspective, the **removal of the human intent from a specific attack weakens moral responsibility by preventing considerations of humanity**. There may be a *causal explanation* for why these combatants were attacked (i.e. they corresponded to the target signature) but we may not be able to offer a *reason*, an ethical justification, for that attack (i.e. why were they attacked in the specific circumstances?). Since the process of reason-giving and justification establishes moral

[43] This does not mean it is necessarily simple, functionally, to identify objects (e.g. vehicles, buildings), since they change status over time (between military objective and civilian object), and objects used by civilians and the military can share similar characteristics.

[44] A Leveringhaus, *Ethics and Autonomous Weapon Systems*, *op. cit.* (footnote 30), 2016, pp. 92–93.

[45] N Melzer, *Interpretive guidance on the notion of direct participation in hostilities under international humanitarian law*, ICRC, Geneva, 2016. Chapter IX: Restraints on the use of force in direct attack, p. 82: "In situations of armed conflict, even the use of force against persons not entitled to protection against direct attack remains subject to legal constraints. In addition to the restraints imposed by international humanitarian law on specific means and methods of warfare, and without prejudice to further restrictions that may arise under other applicable branches of international law, the kind and degree of force which is permissible against persons not entitled to protection against direct attack must not exceed what is actually necessary to accomplish a legitimate military purpose in the prevailing circumstances."

responsibility, and makes people feel they are treated justly, autonomous technology risks blocking this process and diminishing it.

4.2 Transparency in human-machine interaction

Machine control and human control have different strengths and weaknesses. As currently understood, machines have limited decision-making capacities and limited situational awareness but can respond very quickly, and according to specific parameters (although, of course, this is a fast-developing field, especially with respect to artificial intelligence (AI) – *see Section 5.1*). In contrast, humans have a limited attention span and field of perception but global situational awareness of their environment, and sophisticated decision-making capacities**. This difference gives rise to a number of problems in human-machine interaction that are relevant to discussions about autonomous weapon systems**, including: **automation bias** – where humans place too much confidence in the operation of an autonomous machine; **surprises** – where a human is not fully aware of how a machine is functioning at the point s/he needs to take back control; and the "**moral buffer**" – where the human operator shifts moral responsibility and accountability to the machine as a perceived legitimate authority.[46]

This raises additional questions about how moral responsibility and accountability can be ensured in the use of an autonomous weapon system, including whether there will be sufficient transparency in the way it operates, and its interaction with the environment, to be sufficiently understood by humans. To address this concern, a human operator may need to have continuous situational awareness during the operation of an autonomous weapon system, as well as a two-way communication link to receive information and give updated instructions to the system, if necessary, as well as sufficient time to respond or change the course of action, where necessary.

These types of human-machine **problems are already evident in existing civilian autonomous systems**. One example is the accident that resulted when the pilot of a passenger aircraft had to re-take control following a failure in the autopilot system but was not sufficiently aware of the situation to respond in the correct way.[47] Other accidents have happened with car "autopilot" systems, where drivers relied too heavily on a system with limited capacity.[48] And there are also parallels with autonomous financial trading systems, causing so-called "flash crashes" in ways not predictable by human traders overseeing them, and not preventable owing to the extremely short time-scales involved.[49]

[46] M Cummings, "Automation and Accountability in Decision Support System Interface Design", *Journal of Technology Studies*, Vol. XXXII, No. 1, 2006: "… decision support systems that integrate higher levels of automation can possibly allow users to perceive the computer as a legitimate authority, diminish moral agency, and shift accountability to the computer, thus creating a moral buffering effect".

[47] See, for example: R Charette, "Air France Flight 447 Crash Causes in Part Point to Automation Paradox", IEEE Spectrum, 2012: https://spectrum.ieee.org/riskfactor/aerospace/aviation/air-france-flight-447-crash-caused-by-a-combination-of-factors.

[48] J Stewart, "People Keep Confusing Their Teslas for Self-Driving Cars", *Wired*, 25 January 2018: https://www.wired.com/story/tesla-autopilot-crash-dui.

[49] US Securities & Exchange Commission, *Findings regarding the market events of 6 May, 2010. Reports of the staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues, 30 September 2010*.

5. PREDICTABILITY, RELIABILITY AND RISK

Unpredictability and unreliability have been raised as key issues for any legal assessment of autonomous weapon systems,[50] as well as for the risks their use may pose,[51] in particular for civilians. However, these factors are also closely connected to ethical questions of human agency and moral responsibility (*see Sections 3 and 4*).

One way to think about predictability and reliability in autonomous (weapon) systems is as **means of connecting human agency and intent with the eventual outcome and consequences** of the machine's operation. **Predictability** is the ability to "[s]ay or estimate that (a specified thing) will happen in the future or will be a consequence of something".[52] Applied to an autonomous weapon system, predictability is knowledge of how it will likely function in any given circumstances of use, and the effects that will likely result. **Reliability** is "[t]he quality of being trustworthy or performing consistently well".[53] In this context, reliability is knowledge of how consistently the system will function as intended, i.e. without failures or unintended effects.

Degrees of unpredictability and unreliability in the use of an autonomous weapon system might: be inherent to the technical design of the weapon system; arise from the nature of the environment (e.g. 'uncluttered' deep sea versus 'cluttered' populated area); and/or be due to the interaction of the weapon system with the environment. Unpredictability and unreliability in the environment may also vary over time and within a given area (depending on the nature of the environment).

If one recognizes the argument of the necessity for human agency and intent in decisions to use force (*see Section 3*) and the difficulties raised by autonomy for moral responsibility and accountability (*see Section 4*), it follows that the **use of weapon systems that lead to unpredictable and unreliable consequences, and therefore heightened risks for civilians, will accentuate these ethical concerns**. Unpredictability and unreliability, in that sense, are both legally and ethically problematic. However, predictability and reliability, in themselves, do not necessarily resolve ethical questions. For example, an autonomous weapon system might be highly predictable and reliable in attacking combatants, but it could still raise ethical concerns with respect to human agency and human dignity (*see Section 3*).

Of course, there are **only ever degrees of predictability and reliability in complex software-controlled systems**. Unpredictable and unreliable operations may result from a variety of factors, including: **software errors** and **system flaws; human cognitive bias** in dismissing certain possibilities; in-built **algorithmic bias**; [54] "**normal accidents**", where there is no clear error, but a system still does not function as expected; and deliberate **hacking**, **spoofing** or **cyber attacks**.

It is also important to emphasize that nothing is one hundred per cent predictable and reliable, including non-autonomous, human-controlled, weapon systems. Although it is clear that a high degree

[50] N Davison, *Autonomous weapon systems under international humanitarian law*, *op. cit*. (footnote 4), 2017; ICRC, *Views of the ICRC on autonomous weapon systems*, *op. cit.* (footnote 4), 11 April 2016; W Wallach, "Predictability and Lethal Autonomous Weapons Systems (LAWS)", in German Federal Foreign Office, *Lethal Autonomous Weapons Systems: Technology, Definition, Ethics, Law & Security*, 2016, pp. 295–312.

[51] See, for example: P Scharre, Autonomous Weapons and Operational Risk, Center for a New American Security (CNAS), February 2016; UNIDIR, *Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies*, 2016.

[52] *Oxford Dictionary of English*: https://en.oxforddictionaries.com/definition/predictability.

[53] *Ibid*: https://en.oxforddictionaries.com/definition/reliability.

[54] See, for example: A Caliskan, J Bryson and A Narayanan, "Semantics derived automatically from language corpora contain human-like biases", *Science*, Vol. 356, Issue 6334, 2017, pp. 183–186; C O'Neil, *Weapons of Math Destruction: How big data increases inequality and threatens democracy*, Crown, New York, 2016.

would be demanded in safety-critical autonomous systems, such as weapon systems, questions remain about the level of predictability and reliability required to satisfy ethical (and legal) considerations.

5.1 Artificial Intelligence (AI) and unpredictability

For many considering the implications of autonomous weapon systems, the key change in recent years – and **a fundamental challenge for predictability – is the further development of artificial intelligence (AI)**, and **especially AI algorithms that incorporate machine learning**. In general, machine-learning systems can only be understood at a particular moment in time. The "behaviour" of the learning algorithm is determined not only by initial programming (carried out by a human) but also by the process in which the algorithm itself "learns" and develops by "experience". This can be **offline learning by training** (before deployment) and/or **online learning by experience** (after deployment) while carrying out a task.

Deep learning – where an algorithm develops by learning data patterns rather than learning a specific task – further complicates the ability to understand and predict how the algorithm will function, once deployed. It can also add to the problem of biases that can be introduced into an algorithm through limitations in the data sets used to "train" it. Or a learning system may simply have learned in a way that was not intended by the developer.

Complicating matters further, **humans' current ability to interrogate machine-learning algorithms is limited**. Such systems are often described as "back-boxes"; the inputs and outputs may be known but the *process* by which a system converts an input to an output is not known. This type of system can be tested to help determine its functioning in different environments. However, there are significant limits in current abilities to verify the functioning of these systems, a task that becomes harder the more actions there are in the repertoire of the system and the more complex the inputs. If a system continues to learn after being tested, then the verification and validation (checks to determine if a system will operate as intended in a given environment) are no longer meaningful. **This type of autonomous system would be inherently unpredictable** (owing to its technical design) and, if applied to targeting, for example, the link between human intent and eventual outcome would effectively be severed.[55]

Questions about **AI and learning algorithms in weapon systems and targeting functions are no longer theoretical**. As with civilian digital technology, big data are an increasingly important resource, and the focus of data exploitation and analysis efforts is on AI algorithms. For the military, this promises a capability advantage for decision-making in data-rich conflict environments. And despite the risks of unpredictability, which may conflict with military commanders' propensity for command and control, there is significant and increasing interest among the major powers in the military applications of AI,[56] including projects underway to apply machine learning to automatic target recognition and

[55] From a legal perspective, when considering the obligation of States to review new weapons before their deployment and use under Article 36 of Additional Protocol I to the Geneva Conventions, it is difficult to see how a weapon system that autonomously changes its own functioning could ever be approved, since what had been tested and verified at one point in time would not be valid for the future. See: ICRC, *Autonomous weapon systems: Implications of increasing autonomy in the critical functions of weapons*, *op. cit.* (footnote 2), 2016, p. 13.

[56] See, for example: United States Department of Defense, *Summer Study on Autonomy*, Defense Science Board, June 2016; M Cummings, *Artificial Intelligence and the Future of Warfare*, Chatham House, International Security Department and US and the Americas Programme, January 2017; G Allen and T Chan, *Artificial Intelligence and National Security*, Harvard Kennedy School, Belfer Center for Science and International Affairs, 2017; E Kania, *Battlefield Singularity. Artificial Intelligence, Military Revolution, and China's Future Military Power*, Center for a New American Security (CNAS), 2017; "Artificial Intelligence and Chinese Power", Associated Press, 2017; "Putin: Leader in artificial intelligence will rule world", CNBC, 4 September 2017: https://www.cnbc.com/2017/09/04/putin-leader-in-artificial-intelligence-will-rule-world.html.

identification.[57] AI systems may not even need to have a physical component to raise ethical (and legal) questions if their outputs, as "decision aids," are applied to targeting decisions, especially in the absence of cross-checking, or balancing, with other sources of information before human authorization to attack (as over-reliance on algorithmic output would diminish the meaning of the consequent human decision). However, if such AI systems are used directly to control the initiation of an attack by an autonomous weapon system, these concerns would be particularly serious. More broadly, there is growing appreciation of the risks of use, and misuse, of AI across the digital, physical and political domains, and the implications for international security.[58]

The degree of **predictability and reliability of autonomous (weapon) systems affects the trust of humans in that system** – especially in relation to the link between human intention and the eventual "action", or operation, of the system – and this trust is also affected by the degree to which the operation of the system can be explained – or explain itself (e.g. with in-built "explainable AI").[59]

There are now more and more **initiatives addressing these ethical questions for AI systems in general**, including the Institute of Electrical and Electronics Engineers (IEEE)'s Global Initiative on Ethics of Autonomous and Intelligent Systems, which is working on "ethically aligned design" standards for AI and autonomous systems,[60] **and for robotic systems, in particular**.[61] The **Asilomar AI Principles** recently developed by the Future of Life Institute are interesting in this respect. In warning against an AI arms race,[62] they highlight ethical concerns raised by AI systems in general, noting the need for safety, failure transparency, responsibility of developers, alignment with human values and human control over delegation of decisions to AI systems.[63]

5.2 Ethics and risk

Unpredictability and unreliability in autonomous weapon systems **also contribute to the level of risk that the use of the weapon will lead to unacceptable consequences**, in particular for civilians, which raises ethical (as well as legal) issues. Since assessing risk requires an assessment of probability and consequence, machine-learning systems, for example, present immediate problems. Where there is inherent unpredictability in the functioning of a system it may not be possible to assess the *probability*

[57] See, for example: J Keller, DARPA TRACE program using advanced algorithms, embedded computing for radar target recognition", *Military & Aerospace Electronics*, 2015: http://www.militaryaerospace.com/articles/2015/07/hpec-radar-target-recognition.html; D Lewis, N Modirzadeh and G Blum, "The Pentagon's New Algorithmic-Warfare Team", *Lawfare*, 2017: https://www.lawfareblog.com/pentagons-new-algorithmic-warfare-team.

[58] Future of Humanity Institute, University of Oxford; Centre for the Study of Existential Risk, University of Cambridge; Center for a New American Security; Electronic Frontier Foundation; and OpenAI, *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, 2018: https://maliciousaireport.com.

[59] See, for example: DARPA, *Explainable Artificial Intelligence (XAI)*: https://www.darpa.mil/program/explainable-artificial-intelligence.

[60] Institute of Electrical and Electronics Engineers (IEEE), *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*, http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

[61] See, for example: Engineering and Physical Sciences Research Council (EPSRC), *Principles of Robotics*, https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/; J Bryson, "The meaning of the EPSRC principles of robotics", *Connection Science*, Vol. 29 No. 2, 2017, pp. 130–136.

[62] Future of Life Institute, *Asilomar AI Principles*, 2017: https://futureoflife.org/ai-principles/: "18) AI Arms Race: An arms race in lethal autonomous weapons should be avoided."

[63] *Ibid*. "6) Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible. 7) Failure Transparency: If an AI system causes harm, it should be possible to ascertain why. … 9) Responsibility: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications. 10) Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation. 11) Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity. … 16) Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives."

of a certain action, and so determining risk becomes problematic. The introduction of this unpredictability into system design is therefore a significant concern in managing risk. From a purely ethical perspective, some have even argued that creating an unreasonable risk should be considered harm, and ethically wrong, even if that risk does not materialize.[64]

The level of risk also relates to the potential *consequences* of an unpredicted or unintended action, which will also be determined by the specific type of autonomous weapon system and the context of its use, including uses that were not originally foreseen. Some emphasize that the destructive power of the weapon system – in terms of size of munition or potential destructive effects – is an important factor in determining the level of risk, and therefore for an ethical assessment. For example, few would argue for development of autonomous nuclear weapon systems, even if predictability and reliability could be assured as extremely high. However, others are sceptical of a focus on the destructive power, since relatively low-power weapons – such as an autonomous machine-gun system – could still have serious consequences and be used to kill and injure many people *(see also Section 6*). In summary, while predictability and reliability may reduce the risks of unintended consequences in the operation of an autonomous weapon system, they do not, in themselves, eliminate risk.

6. ETHICAL ISSUES IN CONTEXT

Another aspect to consider is whether ethical assessments of autonomous weapon systems vary according to context. In particular, **do specific characteristics of an autonomous weapon system, and the way it is used, have an influence on its ethical acceptability?** For example: the task the weapon is used for, the type of target, the operating environment, the timeframe of operation and the scope of movement over an area.

When discussing different types of autonomous weapon systems, **in different scenarios and contexts, different views tend to emerge on ethical acceptability**. These assessments tend to vary according to the core determinations of human agency in the decision-making process and human dignity (*see Section 3*), associated moral responsibility (*see Section 4*) and, especially, the degree of predictability and risk (*see Section 5*), since contextual factors can have a significant impact on the last of these.

6.1 Constraints in time and space

A longer **timeframe** and/or increased **scope of movement over an area are major factors in contributing to uncertainty** between the point of activation of an autonomous weapon system and the eventual attack that results. As discussed, **an autonomous weapon system** – since it selects and attacks targets independently (after launch or activation) – **creates varying degrees of uncertainty as to exactly when, where and/or why the resulting attack will take place** (*see also Section 4.1*).[65] This is accentuated by wider temporal and spatial boundaries because of greater room for variations in the operational environment over an area, and evolution of that environment over time, both of which may affect the consequences of activation.

Uncertainties introduced by autonomy are clearly a problem from a legal perspective, to the extent that they may prevent the commander or operator from making judgements and taking decisions in line with their legal obligations – of distinction, proportionality and precautions – in carrying out attacks in armed conflict. However, uncertainties also raise concerns from an ethical perspective

[64] C Finkelstein, "Is Risk a Harm?" *University of Pennsylvania Law Review*, No. 263, 2003.

[65] This is in contrast to a long-range non-autonomous weapon system, such as a cruise missile, which may travel long distances, with a significant delay between launch and impact, but is intended to hit a specific target at a specific point in time. (It may also have the capacity to be manually or automatically deactivated after launch.)

because they can decouple human agency and intent in the decision to use force from the eventual consequences, even if the resulting attack is lawful (*see Section 3*).

There are **different dimensions to the issue of temporal constraints**. One is the elapsed time between the point of activation of an autonomous weapon system and the point at which a resulting attack takes place. For example, there is a **significant difference in the level of uncertainty in circumstances that may result during a ten-minute flight time versus a two-day loiter time** (also depending on the operating environment). There are parallels, here, with mine warfare; a major problem with anti-personnel mines, which contributed to their indiscriminate effects and eventual prohibition, was the lack of control over the period during which they could autonomously operate. Once laid by humans, and unless fitted with self-destruct or self-neutralizing features, landmines remain activated indefinitely, and the initial user has no further control over the eventual attack and the nature of the victim.

Mines that stay active indefinitely also raise **another time-related concern: the absence of an "off switch"**. With autonomous weapon systems, the uncertainty over when, where and/or why an attack takes place could be extended indefinitely if there is no capacity to deactivate the system after launch or activation. Unless the system has an automatic self-destruct or self-neutralizing feature (the reliability of which can also vary, as was the case with landmines), the ability to deactivate an autonomous weapon system would require a communication link to a human operator to be retained. Since changes in the operational environment may require deactivation at any point following activation, there is a strong argument for enabling constant human supervision and the ability to intervene and deactivate, as is the case with many existing autonomous weapon systems, such as counter-rocket, artillery and mortar weapons.[66]

A further aspect of the temporal issue is **human reaction time**. Some existing autonomous weapon systems are, by design, intended to initiate an attack quicker than is humanly possible. While speed may create a military advantage – for example, in the case of time-constrained missile and counter-rocket, artillery and mortar defence – it also erodes the potential for human intervention to prevent an unlawful, unnecessary or accidental attack. Even with continuous human supervision, it may only be possible to deactivate a weapon system after a problematic attack in order to prevent further negative consequences, and whether or not this is an acceptable risk may depend on the predictability and reliability of the weapon, the operating environment, as well as the task for which it is used and the target against which it is employed.

6.2 Constraints in operating environments, tasks and targets

The task for which an autonomous weapon system is used and the environment in which it is used can also be significant for ethical assessments. In situations **where there are fewer risks to civilians or civilian objects, some have argued there may also be fewer ethical concerns raised by autonomy** – in terms of reduced human agency. For example, it has been suggested that autonomous deep-sea, anti-submarine warfare and autonomous close-in air defence at sea may be more ethically acceptable, owing to the relatively uncluttered and simple nature of the operating environments, and the reduced numbers of civilians and civilian objects, compared with populated areas on the coast or inland – and, therefore, potentially more predictable, in terms of consequences, and lower-risk.[67]

Further, there is the issue of whether an autonomous weapon system is used for defensive or offensive tasks. Some suggest there may be an ethical distinction between a "defensive" weapon system – such

[66] Such a requirement could limit the utility of autonomous weapon systems where constant communication is not feasible, such as underwater.
[67] R Sparrow and G Lucas, "When Robots Rule the Waves?" *Naval War College Review*, 69(4), 2016, pp. 49–78.

as a missile or counter-rocket, artillery and mortar defence weapon, or a "sentry" weapon guarding a border – and an "offensive" system, which actively searches for targets. However, others caution that the distinction between "offensive" and "defensive" is not clear operationally (and legally, the same rules apply to the use of force or conduct of hostilities), and that a weapon system introduced for a "defensive" task may later be used in an "offensive" role.

Perhaps **the most significant contextual factor that gives rise to ethical concerns, however, is the nature of the target**, and whether the weapon system only targets objects or attacks humans directly. The fundamental anxiety in the ethical discourse is about anti-personnel autonomous weapon systems, especially, it is argued, with respect to: lack of human agency and intent in decisions to use force; the loss of human dignity on the part of those combatants targeted,[68] and of civilians that are put at risk as a consequence of legitimate attacks on military targets; and the implications for moral responsibility (*see Sections 3 and 4*).

7. PUBLIC AND MILITARY PERCEPTIONS

Although public opinion does not necessarily equal public conscience, and ethics, as a formal mode of criticism, should not be reduced to opinion polls, it is useful to explore the perspectives on autonomous weapon systems from different constituents of society – including the public, the military, and the scientific and technical communities.[69]

Public opinion may not provide evidence-based answers to ethical questions, especially when those surveyed have different understandings of the questions and the concept of an autonomous weapon system. **However, opinion polls can spark debate** and illustrate a significant interest in and engagement with the topic by different constituents, **as well as revealing trends related to public-conscience concerns**.

7.1 Opinion surveys

There have been several surveys of public opinion in this field.[70] **Many have contrasted remote-controlled armed drones with autonomous weapon systems**, in order to differentiate reactions to autonomy specifically from robotic-weapons platforms in general. In 2011, Moon, Danielson and Van der Loos found greater rejection of autonomous weapon systems (81% against, 10% in favour) than of remote-controlled drones (53% against, 35% in favour) based on three major rationales: preservation of human responsibility and accountability; scepticism about the technology, and therefore risks for civilians; and assertions that humans should always make life-or-death decisions.[71]

In 2015, an Open Roboethics Initiative survey gathered the views of 1000 people from 49 different countries. It, too, found a significant rejection of autonomous weapon systems (67% said all types

[68] R Sparrow, "Twenty seconds to comply: Autonomous Weapon Systems and the recognition of surrender", *International Law Studies*, 91, 2015, pp. 699–728.
[69] R Sparrow, "Ethics as a source of law: The Martens clause and autonomous weapons", *ICRC Humanitarian Law & Policy Blog*, 14 November 2017: http://blogs.icrc.org/law-and-policy/2017/11/14/ethics-source-law-martens-clause-autonomous-weapons.
[70] Including: L Moshkina and R Arkin, "*Lethality and Autonomous Systems: The Roboticist Demographic",* IEEE International Symposium on Technology and Society, 2008; Prof. C Carpenter, *US public opinion on autonomous weapons*, University of Massachusetts Department of Political Science, 2013; M Horowitz, "Public opinion and the politics of the killer robots debate", *Research and Politics*, January–March 2016.
[71] A Moon, P Danielson and M Van der Loos, "Survey-based Discussions on Morally Contentious Applications of Interactive Robotics", *International Journal of Social Robotics*, Volume 4, Issue 1, 2012, pp 77–96.

should be banned) and stronger views based on the type of task (85% should not be used for "offensive purposes"). The rejection of autonomous weapons was also greater in comparison with remote-controlled weapons (71% would prefer their military to use remote-controlled weapons in warfare; 60% would prefer to be attacked by remote-controlled rather than autonomous weapons).[72]

A 2017 **IPSOS poll of 11,500 respondents in 25 countries also found overall opposition** to autonomous weapon systems (56% against, 24% in favour), although the poll also revealed regional variations, with the greatest opposition in Russia (69% against), Peru (67% against), Spain (66% against) and Argentina (66% against), and the least in India (31%), China (36%) and the United States (45%).[73]

While each study has its limitations, these polls reflect trends that are worth exploring further. Why do people tend to prefer attacks to be carried out by remote-controlled rather than autonomous weapon systems? How much significance is placed on reservations about the technology and its consequences, and how much on ethical concerns about human agency, human dignity and the view that machines must not take decisions on the use of force?

7.2 Contrasting military and public perceptions

Another 2017 survey contrasted perceptions of remote-controlled armed drones and autonomous weapon systems among the public in the United States, and civilian and military personnel of the Dutch Ministry of Defence.[74] The Ministry of Defence personnel had less trust, confidence and support for the "actions" taken by autonomous weapon systems compared with remote-controlled systems but considered them equally "fair". Respondents were, generally, more anxious about the consequences of using autonomous weapon systems, and concern about a lack of respect for human dignity was one of the main objections, when compared with human-operated drones resulting in the same consequences. **In comparisons between military and public perceptions, most notable was the similar level of concern about a loss of human dignity**, which may indicate some common ground among different constituents.

8. CONCLUSIONS

Ethics, humanity and the dictates of the public conscience are at the heart of the debate about the acceptability of autonomous weapon systems. From the ICRC's perspective, ethics provides another avenue – alongside legal assessments and technical considerations – to help determine the necessary type and degree of human control that must be retained over weapon systems, and the use of force, and to elucidate where States must establish limits on autonomy in weapon systems.

Considerations of humanity and the public conscience provide **ethical guidance for discussions**, and there is a requirement to connect them to legal assessments via the **Martens Clause – a safety net for humanity**. These ethical considerations go beyond whether autonomous weapon systems are compatible with our *laws* to include fundamental questions of whether they are acceptable to our *values*. And such debates necessarily require the engagement of various constituents of society.

[72] Open Roboethics Initiative, *The Ethics and Governance of Lethal Autonomous Weapons Systems: An International Public Opinion Poll*, 9 November 2015.
[73] IPSOS, Three in ten Americans support using Autonomous Weapons, 7 February 2017.
[74] I Verdiesen, *Agency perception and moral values related to Autonomous Weapons: An empirical study using the Value-Sensitive Design approach*, Masters of Science, Faculty of Technology, Policy and Management, TU Delft, 2017.

Several ethical issues appear central to establishing constraints on autonomy in weapon systems. Perhaps the most powerful ethical concerns are those that transcend context – whether during armed conflict or in peacetime – and transcend technology – whether simple or sophisticated.[75] These are concerns about **loss of human agency in decisions to use force** – decisions to kill, injure and destroy – **loss of human dignity in the process of using force**, and **erosion of moral responsibility for these decisions.**

The importance of **retaining human agency – and intent – in these decisions** is one of the central ethical arguments for limits on autonomy in weapon systems. Many take the view that decisions to kill, injure and destroy must not be delegated to machines, and that humans must be present in this decision-making process sufficiently to preserve a direct link between the intention of the human and the eventual operation of the weapon system. It is not enough simply to say that "humans have developed, deployed and activated the weapon system". **There must be a direct connection between the human rationale for activation of an autonomous weapon system in the specific circumstances and the consequences of the resulting attack**. But questions remain about how close this connection must be, and what form it must take.

Human dignity is another core ethical consideration that is linked to concerns about loss of human agency. The central argument is that it matters not just *if* a person is killed or injured but *how* they are killed or injured, and the **process by which these decisions are made is as important as the results**. If human agency is lacking to the extent that machines have effectively, and functionally, been delegated these decisions, then, according to this argument, it undermines the human dignity of those combatants targeted, and of civilians that are put at risk as a consequence of legitimate attacks on military targets. If human agency is retained, on the other hand, it is an acknowledgement of humanity in that decision to use force and the resulting consequences.

The need for human agency is also linked to **moral responsibility and accountability** for decisions to use force. These are human responsibilities (both ethical and legal), which **cannot be transferred to inanimate machines, or computer algorithms**, since it is humans that have both rights and responsibilities in relation to these decisions. From an ethical perspective, it is not sufficient only to assign legal responsibility to a commander or operator who activates an autonomous weapon system. Humans must uphold their **moral responsibility**, requiring not only a causal explanation but also a justification for the resulting use of force. Autonomous weapon systems complicate this justification because of the more generalized nature of the targeting decisions, which risks eroding – or diffusing – moral responsibility.

Predictability and **reliability** in using an autonomous weapon system are ways of connecting human agency and intent to the eventual consequences of the resulting attack. **A lack of predictability**, whether inherent to the weapon system design or due to interaction with the environment, **raises serious ethical (and legal) concerns owing to a lack of foreseeability of the consequences** and associated risks, in particular for civilians.

As weapons that self-initiate attacks, **autonomous weapon systems all raise questions about predictability**, owing to varying degrees of uncertainty as to exactly when, where and/or why a resulting attack will take place. However, the **application of AI and, in particular, machine learning, to targeting functions accentuates this problem, raising fundamental questions of inherent**

[75] Although there are different views among experts on the issue of technology. Some make a distinction between "automated" and "autonomous" weapons and focus their concerns on systems controlled by complex AI algorithms rather than simpler software. Others, including the ICRC, note the lack of a clear technical distinction between the two, and argue that "all such weapons raise the same core legal and ethical questions". See: ICRC, *Autonomous weapon systems: Implications of increasing autonomy in the critical functions of weapons*, *op. cit.* (footnote 2), 2016, p. 8.

unpredictability by design and heightening concerns about the loss of human agency, moral responsibility and human dignity.

Context also affects ethical assessments of autonomous weapon systems, owing to the impact on the predictability of the outcomes of their use, the nature of the consequences and the overall level of risk that results. Constraints on the **timeframe of operation** and **scope of movement over an area** are key factors, as are the **task** for which the weapon is used and the **operating environment** in which it is activated.

However, from an ethical perspective, perhaps the most important contextual factor is the type of target. **Core concerns about human agency, human dignity and moral responsibility are most acute in relation to the notion of anti-personnel autonomous weapon systems that target humans directly**. These concerns may be one reason – together with legal considerations and technical limitations – why the use of autonomous weapon systems to date has been constrained to anti-materiel systems,[76] targeting projectiles, vehicles, aircraft or other objects, even if these systems pose dangers to humans inside or in proximity to objects.[77]

8.1 An ethical basis for human control?

From the ICRC's perspective**, ethical considerations very much parallel the requirement for a minimum level of human control over weapon systems and the use of force**, to ensure compliance with international legal obligations that govern the use of force in armed conflict and in peacetime.[78]

From an ethical viewpoint, "**meaningful", "effective" or "appropriate" human control would be the type and degree of control that preserves human agency and upholds moral responsibility in decisions to use force**. This does not necessarily exclude autonomy in weapon systems, but it requires a sufficiently direct and close connection to be maintained between the human intent of the user and the eventual consequences of the operation of the weapon system in a specific attack. This, in turn, will necessitate limits on autonomy.

Ethical and legal considerations may demand some similar constraints on autonomy in weapon systems so that meaningful human control is maintained – in particular, with respect to: human supervision and the ability to intervene and deactivate; technical requirements for predictability and reliability (including in the algorithms used); and operational constraints on the task for which the weapon is used, the type of target, the operating environment, the timeframe of operation and the scope of movement over an area.[79]

However, the **combined and interconnected ethical concerns** about loss of human agency in decisions to use force, diffusion of moral responsibility and loss of human dignity **could have the most far-reaching consequences, perhaps precluding the development and use of anti-personnel autonomous weapon systems, and even limiting the applications of anti-materiel systems**, depending on the risks that destroying materiel targets present for human life.

---

[76] There have been reports that some anti-personnel "sentry" weapon systems have autonomous modes. However, as far as is known to the ICRC, "sentry" weapon systems that have been deployed still require human remote authorization to launch an attack (even though they may identify targets autonomously). See also footnote 3.

[77] Including through accidents. See, for example, "fratricide" incidents discussed in: J Hawley, *Automation and the Patriot Air and Missile Defense System*, *op. cit.* (footnote 24)*,* 2017.

[78] N Davison, *Autonomous weapon systems under international humanitarian law*, *op. cit.* (footnote 4), 2017; M Brehm, *Defending the Boundary: Constraints and Requirements on the Use of Autonomous Weapon Systems Under International Humanitarian and Human Rights Law*, Geneva Academy Briefing no. 9, 1 May 2017.

[79] ICRC, *Statement to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on "Lethal Autonomous Weapon Systems"*, *op. cit.* (footnote 4)*,* 15 November 2017.