Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects

28 August 2018

English only

Second Session
Geneva, 27 - 31 August 2018
Item 6 of the provisional agenda
Other matters

Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems

Submitted by the United States

Ensuring that Machines Effectuate Human Intent in Using Force

1. In our view, the key issue for human-machine interaction in emerging technologies in the area of LAWS is ensuring that machines help effectuate the intention of commanders and the operators of weapons systems. This is done by, *inter alia*, taking practical steps to reduce the risk of unintended engagements and to enable personnel to exercise appropriate levels of human judgment over the use of force.

2. This approach supports compliance with the law of war. Weapons that do what commanders and operators intend can effectuate their intentions to conduct operations in compliance with the law of war and to minimize harm to civilians and civilian objects.

3. This paper discusses a number of measures the United States is taking to ensure that new weapons help effectuate the commander's intent. These measures and policies are set forth in U.S. Department of Defense Directive 3000.09, Autonomy in Weapon Systems (DoD Directive 3000.09). DoD Directive 3000.09 was initially issued in 2012 after a DoD working group considered DoD's past practice in using autonomy in weapon systems, including lessons learned, and potential future applications of autonomy in weapon systems.

Minimizing unintended engagements

4. DoD Directive 3000.09 states that one of its purposes is to establish "guidelines designed to minimize the probability and consequences of failures in autonomous and semi-autonomous weapon systems that could lead to unintended engagements."[1]

5. DoD Directive 3000.09 defines "unintended engagement" as "[t]he use of force resulting in damage to persons or objects that human operators did not intend to be the

[1] DoD Directive 3000.09, 1.a.

targets of U.S. military operations, including unacceptable levels of collateral damage beyond those consistent with the law of war, ROE, and commander's intent."[2]

6. For example, accidental attacks that killed civilians or friendly forces would be "unintended engagements" under DoD Directive 3000.09.

7. Similarly, even an attack against authorized targets could be "unintended" if there are significant changes to the factual context between the time of authorization and the engagement (for example, if a cease-fire agreement is negotiated). In this regard, DoD Directive 3000.09 requires that autonomous and semi-autonomous weapon systems be designed to "[c]omplete engagements in a timeframe consistent with commander and operator intentions and, if unable to do so, to terminate engagements or seek additional human operator input before continuing the engagement."[3]

Ensuring appropriate levels of human judgment over the use of force

8. DoD Directive 3000.09 requires that autonomous and semi-autonomous weapon systems "be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force."[4]

9. "Appropriate" is a flexible term that reflects the fact that there is not a fixed, one-size-fits-all level of human judgment that should be applied to every context. What is "appropriate" can differ across weapon systems, domains of warfare, types of warfare, operational contexts, and even across different functions in a weapon system. Some functions might be better performed by a computer than a human being, while other functions should be performed by humans.

10. In some cases, less human involvement might be more appropriate. For example, in certain defensive autonomous weapon systems, such as the Phalanx Close-In Weapon System, the AEGIS Weapon System, and Patriot Air and Missile Defense System, the weapon system has autonomous functions that assist in targeting incoming missiles or other projectiles. The machine can strike incoming projectiles with much greater speed and accuracy than a human gunner could achieve manually. As weapons engineers improve the effectiveness of autonomous functions, more situations will likely arise in which the use of autonomous functions is more appropriate than manual control.

11. "Human judgment over the use of force" is distinct from human control over the weapon. For example, an operator might be able to exercise meaningful control over every aspect of a weapon system, but if the operator is only reflexively pressing a button to approve strikes recommended by the weapon system, the operator would be exercising little, if any, judgment over the use of force. On the other hand, judgment can be implemented through the use of automation. For example, the extensive automation of functions in a weapon system could allow the operator to exercise better judgment over the use of force by removing the need to focus on basic tasks and to give him or her more time to understand the broader situation. Similarly, the use of algorithms or even autonomous functions that take control away from human operators can better effect human intentions and avoid accidents. A useful case to consider may be the Automatic Ground Collision Avoidance System developed by the U.S. Air Force that has helped prevent so-called "controlled flight into terrain" accidents. The system assumes control of the aircraft when an imminent collision with the ground is detected and returns control back to the pilot when the collision is averted.

12. DoD Directive 3000.09's requirements that weapons be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force reflect a deliberate decision to permit weapons that are programmed to make "decisions" that relate to targeting.

[2] DoD Directive 3000.09, Glossary.
[3] DoD Directive 3000.09, 4.a.(1)(b); *see also* DoD Directive 3000.09, Enclosure 3, 1.a.(2).
[4] DoD Directive 3000.09, 4.a.

13. Autonomy has already been used sensibly in targeting-related functions such as identifying, selecting, and determining whether and when to engage targets. As we noted in a working paper submitted in 2017, there is no requirement that the machine itself be programmed to make law of war assessments, such as whether the target is a military objective. Rather, there are a variety of ways to ensure that even relatively simple forms of automation can be used appropriately in military operations.

14. For example, an autonomous system might be programmed to operate only within certain geographic boundaries. If deployed and limited to an area that was a military objective, such as an enemy military headquarters complex, then its use would be analogous to the use of other weapons, like artillery, that are used to target areas of land that qualify as military objectives.

15. Similarly, an autonomous system might be equipped with sensors that are designed to detect specific "signatures" – unique, identifying characteristics that would be specific to a military objective, such as frequencies of electromagnetic radiation that are generally not found naturally or among civilian objects. Many States have used weapons that detect the specific electromagnetic signals emitted by enemy radar.

Practical measures to ensure the use of autonomy in weapon system effectuates human intentions

16. DoD Directive 3000.09 establishes a number of requirements – at different stages of the weapon design, development, and deployment process – intended to ensure the use of autonomy in weapon systems effectuates human intentions.

17. A key theme among these requirements is ensuring that systems "function as anticipated."[5] This entails engineering weapon systems to perform reliably, training personnel to understand the systems, and establishing clear human-machine interfaces.

18. First, a variety of measures are taken to ensure that weapons are engineered to perform as expected.

19. DoD Directive 3000.09 establishes requirements for verification and validation and test and evaluation. Before fielding systems that would use autonomy in novel ways, such reviews must "assess system performance, capability, reliability, effectiveness, and suitability under realistic conditions, including possible adversary actions, consistent with the potential consequences of an unintended engagement or loss of control of the system."[6] Such testing should include "analysis of unanticipated emergent behavior resulting from the effects of complex operational environments on autonomous or semi-autonomous systems."[7]

20. DoD Directive 3000.09 also requires that "safeties, anti-tamper mechanisms, and information assurance" have been implemented in autonomous and semi-autonomous weapon systems.[8] These measures are intended to "minimize the probability or consequences of failures that could lead to unintended engagements or to loss of control of the system" by, for example, safeguarding against attempts by unauthorized individuals to fire the weapon.[9]

21. Second, DoD Directive 3000.09 seeks to ensure that personnel properly understand the weapon systems. A key insight from past studies of accidents involving human use of automation, such as studies of accidental shoot-downs of friendly aircraft by the Patriot missile system, is that failures can often result from operator error and that better training and adherence to established tactics, techniques, and procedures (TTPs) and doctrine could prevent mistakes that would result in unintended engagements.

[5] DoD Directive 3000.09, 4.a.(1)(a).
[6] DoD Directive 3000.09, Enclosure 3, 1.b.(3).
[7] DoD Directive 3000.09, Enclosure 2, a.
[8] DoD Directive 3000.09, 4.a.(2)(a).
[9] DoD Directive 3000.09, Enclosure 3, 1.b.(2).

22. Therefore, DoD Directive 3000.09 generally requires the establishment of "[t]raining, doctrine, and tactics, techniques, and procedures."[10] Moreover, before systems that employ autonomy in new ways are fielded, senior officials must determine that "[a]dequate training, TTPs, and doctrine are available, periodically reviewed, and used by system operators and commanders to understand the functioning, capabilities, and limitations of the system's autonomy in realistic operational conditions."[11]

23. Officials responsible for training and equipping forces are to "[c]ertify that operators of autonomous and semi-autonomous weapon systems have been trained in system capabilities, doctrine, and TTPs in order to exercise appropriate levels of human judgment in the use of force and employ systems with appropriate care and in accordance with the law of war, applicable treaties, weapon system safety rules, and applicable ROE."[12]

24. In addition, commanders must use weapons "in a manner consistent with their design, testing, certification, operator training, doctrine, TTPs, and approval as autonomous or semi-autonomous systems."[13]

25. Third, DoD Directive 3000.09 requires that the interface between humans and machines be clear "[i]n order for operators to make informed and appropriate decisions in engaging targets."[14]

26. In particular, DoD Directive 3000.09 requires that "the interface between people and machines for autonomous and semi-autonomous weapon systems shall:

(a) Be readily understandable to trained operators;

(b) Provide traceable feedback on system status;

(c) Provide clear procedures for trained operators to activate and deactivate system functions."[15]

Holistic, Proactive, Review Processes Guided by the Fundamental Principles of the Law of War

27. Emerging technologies are difficult to regulate because technologies continue to change as scientists and engineers develop advancements. A best practice today might not be a best practice in the near future. Similarly, a weapon system that, if built today, would risk creating indiscriminate effects, might, if built with future technologies, prove more discriminating than existing alternatives by reducing the risk of civilian casualties.

28. Thus, rather than seeking to codify best practices or set new international standards, States should seek to exchange practice and implement holistic, proactive review processes that, are guided by the fundamental principles of the law of war.

Holistic processes across the touch points in the human-machine interface

29. The Chair of the GGE has helpfully framed "four broad areas of touch points in the human-machine interface" – 1) "Research & Development"; 2) "Testing and Evaluation," "Verification and Validation," and "Reviews"; 3) "Deployment, Command & Control"; and 4) "Use & Abort."[16]

30. In addressing issues in human-machine interaction, we recommend a holistic approach that considers all the touch points of human-machine interaction. For example,

[10] DoD Directive 3000.09, 4.a.(1).
[11] DoD Directive 3000.09, Enclosure 3, 1.b.(4).
[12] DoD Directive 3000.09, Enclosure 4, 8.a.(5).
[13] DoD Directive 3000.09, Enclosure 4, 10.a.
[14] DoD Directive 3000.09, 4.a.(3).
[15] DoD Directive 3000.09, 4.a.(3).
[16] Chair's summary of the discussion on Agenda item 6(a) 9 and 10 April 2018, Agenda item 6(c) 12 April 2018, Agenda item 6(d) 13 April 2018.

the solution to a problem identified during use of a weapon might be generated by a research laboratory, or an issue identified in the development of a weapon might be resolved by new policies or rules of engagement.

31. As a case in point, trust and accountability issues are posed by the fact that current AI systems often use processes that are opaque to the human operators of the systems. To help address trust and accountability issues, the Defense Advanced Research Projects Agency's Explainable AI project seeks to develop new machine-learning systems that "have the ability to explain their rationale, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future."[17] By seeking to develop AI systems that are more transparent to human operators, such work in the research and development area can address concerns that might be posed by the use of such technology.

Proactive reviews during development and before fielding

32. We also recommend a proactive approach in addressing issues in human-machine interaction. States seeking to develop new uses for autonomy in their weapons should be affirmatively seeking to identify and address these issues in their respective processes for managing the life cycle of such weapons. For example, DoD Directive 3000.09 requires senior officials to review weapon systems that use autonomy in new ways. Such reviews, which are required before a system enters formal development and, again, before fielding, ensure that military, acquisition, legal, and policy expertise is brought to bear before new types of weapons systems are used.

33. This practice in conducting a special policy review is consistent with broader DoD practice in conducting legal reviews of the intended acquisition or procurement of any weapon by the Department of Defense, as reflected in U.S. Department of Defense Directive 5000.01, The Defense Acquisition System. Such reviews, among other things, help ensure consistency with the law of war.

Guidance from the fundamental principles of the law of war

34. In applying holistic approaches and proactive review processes, States should be guided by the fundamental principles of the law of war.

35. The U.S. military has long used the fundamental principles of law of war as a general guide for conduct during war, when no more specific rule applies.[18] These principles are: military necessity, humanity, distinction, proportionality, and honor.[19]

36. These principles have also been the basis for many codifications of the law of war, including the Geneva Conventions of 1949, which, as the International Court of Justice

[17] David Gunning, *Explainable Artificial Intelligence (XAI)*, available at: https://www.darpa.mil/program/explainable-artificial-intelligence.

[18] *See, e.g.*, U.S. Department of Defense Law of War Manual § 2.1.2.2 (June 2015, Updated December 2016) ("When no specific rule applies, the principles of the law of war form the general guide for conduct during war."). U.S. War Department, Part Two, Rules of Land Warfare, Basic Field Manual, Volume VII, Military Law, p.1, ¶4, Jan. 2, 1934 ("Among the so-called unwritten rules or laws of war are three interdependent basic principles that underlie all of the other rules or laws of civilized warfare, both written and unwritten, and form the general guide for conduct where no more specific rule applies, … ."); Instructions for the Government of Armies of the United States in the Field, Prepared by Francis Lieber, Issued as General Orders No. 100, Adjutant General's Office, 1863, arts. 14-16 (discussing the principle of military necessity), art 30 ("No conventional restriction of the modes adopted to injure the enemy is any longer admitted; but the law of war imposes many limitations and restrictions on principles of justice, faith, and honor.").

[19] U.S. Department of Defense Law of War Manual, Chapter II (June 2015, Updated December 2016).

(ICJ) has observed, "are in some respects a development, and in other respects no more than the expression, of" fundamental general principles of international humanitarian law.[20]

37. The practice of resorting to the fundamental principles of the law of war even though specific rules might not apply, has itself been codified in the so-called "Martens Clause." First included in the Preamble to the 1899 Hague Convention II with Respect to the Laws and Customs of War on Land, the clause also is included in a common article to the 1949 Geneva Conventions, which states that denunciation of the Convention "shall in no way impair the obligations which the Parties to the conflict shall remain bound to fulfil by virtue of the principles of the law of nations, as they result from the usages established among civilized peoples, from the laws of humanity and the dictates of the public conscience."[21]

38. The ICJ has observed that, in relation to "the cardinal principles constituting the fabric of humanitarian law," the Martens Clause "has proved to be an effective means of addressing the rapid evolution of military technology."[22] The ICJ's observation has been reflected in the practice of the United States. For example, careful consideration of the principles of military necessity and humanity has been critical to the U.S. Department of Defense's review of the legality of new weapons.[23]

39. In addition to helping to assess whether a new weapon falls under a legal prohibition, the fundamental principles of the law of war may also serve as a guide in answering novel ethical or policy questions in human-machine interaction presented by emerging technologies in the area of LAWS.

40. For example, if the use of a new technology advances the universal values inherent in the law of war, such as the protection of civilians, then the development or use of this technology is likely to be more ethical than refraining from such use.

41. The following questions might be useful to consider in assessing whether to develop or deploy an emerging technology in the area of lethal autonomous weapons systems:

(a) Does military necessity justify developing or using this new technology?

(b) Under the principle of humanity, does the use of this new technology reduce unnecessary suffering?

(c) Are there ways this new technology can enhance the ability to distinguish between civilians and combatants?

(d) Under the principle of proportionality, has sufficient care been taken to avoid creating unreasonable or excessive incidental effects?

(e) Under the principle of the honor, does the use of this technology respect and avoid undermining the existing law of war rules?

"Human Control"

42. The key issue for human-machine interaction in the development, deployment, and use of emerging technologies in the area of lethal autonomous weapons systems is ensuring

[20] Military and Paramilitary Activities in and against Nicaragua, (Nicaragua v. United States of America), Merits, Judgment, I.C.J. Reports 1986, p.14, 113 (June 27, 1986, ¶218).

[21] Geneva Convention for the Amelioration of the Wounded and Sick in Armed Forces in the Field of August 12, 1949, art. 63, 1950 UNTS 32, 68; Geneva Convention for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of the Armed Forces at Sea of August 12, 1949, art. 62, 1950 UNTS 86, 120; Geneva Convention Relative to the Treatment of Prisoners of War of August 12, 1949, art. 142, 1950 UNTS 136, 242; Geneva Convention Relative to the Protection of Civilian Persons in Time of War of August 12, 1949, art. 158, 1950 UNTS 288, 392.

[22] Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, I.C.J. Reports 1996, p. 226, 257 (July 8, 1996, 78).

[23] U.S. Department of Defense Law of War Manual, §§ 6.6.2, 6.6.3.1 (June 2015, Updated December 2016) (discussing the application of the principles of humanity and military necessity in the context of applying the prohibition against weapons calculated to cause superfluous injury).

that when it is necessary to use force, such force is used to effectuate the intentions of commanders and operators. In particular, practical measures should be taken to reduce the risk of unintended engagements (e.g., those resulting from accidents or sabotage) and to ensure that personnel exercise appropriate levels of human judgment over any use of force.

43. We view this as distinct from the concept of "human control," a term that risks obscuring the genuine challenges in human-machine interaction.

44. Practical measures to facilitate effective human-machine interaction – ensuring that force is used to effectuate human intentions – are set forth in DoD Directive 3000.09, Autonomy in Weapon Systems.

45. Seeking to codify best practices or set new international standards for human-machine interaction in this area is impractical because rapid technological advancements may render such practices or standards obsolete shortly after they are established. Instead, States should ensure responsible use of emerging technologies in military operations by implementing holistic, proactive review processes that are guided by the fundamental principles of the law of war.

Terminologies and Conceptualizations: The Misplaced Focus of "Human Control"

46. During the April 2018 session of the GGE, delegations presented a range of different terminologies and conceptualizations regarding human-machine interaction, including human control, supervision, oversight, and judgment. Some have advocated that CCW GGE discussions focus in particular on the issue of "human control" of weapons systems and have advocated for the promulgation of new standards to ensure minimum levels of control or "meaningful human control." The concept of "human control" is subject to divergent interpretations that can hinder meaningful discussion.

47. As we explain below, we believe that emphasis on "control" would obscure rather than clarify the genuine challenges in this area.

International discussions about weapon control systems related to emerging technologies are not likely to produce useful common understandings with respect to all weapons that use such technologies

48. On a practical level, discussions of the technical systems that are used to control weapon systems manually are not likely to advance our collective understanding of the challenges and benefits presented by emerging technologies. How a weapon system is controlled is often very specific to the weapon system, and control systems can vary greatly from system to system. Accordingly, any insight that can be gained from discussing human control of one weapon system may only be of limited relevance to other weapons systems.

49. Similarly, past regulation of weapons systems under international humanitarian law has not included broadly applicable standards for weapon control systems. Moreover, existing international humanitarian law instruments, such as the CCW and its Protocols, do not seek to enhance "human control" as such. Rather, these instruments seek, *inter alia*, to ensure the use of weapons consistent with the fundamental principles of distinction and proportionality, and the obligation to take feasible precaution for the protection of the civilian population. Although control over weapon systems can be a useful means in implementing these principles, "control" is not, and should not be, an end in itself.

Autonomous functions in a weapon system can enhance human control over the use of force

50. Some may think it important to emphasize "human control" because they view developments in the use of automation or autonomy in a weapon system as decreasing human control over the use of force. We believe such a view would be mistaken.

51. Technical sophistication in a weapon system that enables it to perform functions autonomously – what are often called "smart" weapons – does not necessarily mean that

there is any less human involvement in the decision-making of how that weapon is used. The use of technology, such as sensors and computers, allows personnel to set the parameters for when, where, and how force is deployed without manually controlling the weapons system at all times.

52. The use of "smart" weaponry with autonomous functions has increased the degree of control that States exercise over the use of force. For example, many States employ weapons such as Hellfire or Javelin missiles, which use autonomy in critical functions to home-in on targets identified by human operators. Other common weapons, such as the High-speed Anti-Radiation Missile (HARM) or SMArt 155 artillery shells, have autonomous functions that allow them to sense categories of targets according to how they have been programmed and to guide themselves to those targets.

53. Personnel use these weapons with the intention to achieve specific military effects. The fact that the projectile might also "select" a target that has been identified by a human operator or that has been programmed into it and autonomously maneuver itself toward a target does not amount to a delegation of decision-making from humans to machines. Rather, the machine's programming and sensors enable it to effectuate the intentions of the forces using this weapon in a way that is superior to weapons without such programming and sensors.

Manual control of a weapons system is not a prerequisite for holding humans accountable

54. Some may argue that it is important to emphasize control because of concerns that the use of autonomous weapons systems somehow removes individuals from responsibility. However, personnel are responsible for their decisions to use force regardless of the nature of the weapon system they utilize. The lack of a manual control over a weapon system does not remove this responsibility or result in an accountability gap.

55. Computers can enable machines to respond to inputs from sensors through an application of the algorithms or other processes with which they have been programmed. Machines, however, are not intervening moral agents, and human beings do not escape responsibility for their decisions by using a weapon with autonomous functions.

56. When using weapons systems with autonomous functions, the commander must make the legal judgments required by IHL, including by the principles of distinction and proportionality. The human operators of the system and their superior commanders are responsible and accountable for their use of the system, even if that system has sophisticated autonomous functions.