A human touch: autonomous weapons, directive 3000.09, and the "appropriate levels of human judgment over the use of force"

Saxon, D.R.

A Human Touch

Autonomous Weapons, Directive 3000.09, and the "Appropriate Levels of Human Judgment over the Use of Force"

Dan Saxon

The application of International Humanitarian Law (IHL) is an attempt to achieve an equitable balance between humanitarian requirements and the necessities of war.[1] The efforts of IHL and International Human Rights Law (IHRL) to promote the 'humanization of war' presuppose that war's decision-makers are human.[2, 3] Increasingly, however, important decisions of war will be made—directly or indirectly—by machines and virtual networks linking machines which, to varying degrees, are controlled by humans. With advances in artificial intelligence, machines become less dependent on human control and humans become more dependent on machines to take action for them. But how much independence to kill should humans cede to autonomous weapons?[4]

This is not an abstract or merely academic issue. The U.S. government, for example, has begun to develop formal—albeit vague—policies concerning the development and use of semi-autonomous and autonomous weapons.[5] In Department of Defense Directive Number 3000.09, Deputy Secretary of Defense for Policy Ashton B. Carter defines an autonomous weapon system as one that, "once activated, can select and engage targets without further intervention by a

Dan Saxon is an Assistant Professor of International Law at Leiden University College in The Hague. Saxon has served as the legal adviser to the United Nations Commission of Inquiry for Syria.

human operator."[6] This article adopts that relatively broad view for the purpose of the discussion below.

The Directive describes who shall be responsible for, *inter alia*, the lawful design of semi-autonomous and autonomous weapons, their experimentation strategies, human-machine interfaces, operational standards, doctrine, training, hardware and software safety mechanisms, and employment against adversaries.[7, 8] Furthermore, it confirms that persons who authorize, direct, or employ semi-autonomous or autonomous weapons systems must do so with appropriate care. This care must ultimately be consistent with IHL, as well as applicable treaties, weapons system safety rules, and rules of engagement (ROE).[9] Furthermore, the Directive anticipates that "unintended engagements" (i.e. the death and injury of civilians) will occur and obliges military and civilian leaders to design semi-autonomous or autonomous weapons systems so as to minimize the probability of such failures or of "loss of control of the system."[10]

Many of these systems are based on technology exported from democratic states with constitutional commitments to human rights.

DOD Directive 3000.09 affirms that it is U.S. Department of Defense policy that autonomous and semi-autonomous weapons systems shall be designed to allow commanders and operators to exercise "appropriate levels of human judgment over the use of force." The Directive initially limits the use of autonomous weapon systems to the application of non-lethal, non-kinetic force.[11] However, there is a loophole to this restriction. Autonomous weapon systems that operate differently—those which might apply lethal force—may be developed and deployed with approval from the Under Secretary of Defense for Policy; the Under Secretary of Defense for Acquisition, Technology, and Logistics; and the Chairman of the Joint Chiefs of Staff.[12] Before making the decision to develop or deploy a lethal, kinetic autonomous weapon system, these three officials must ensure that the system and components incorporate the necessary capabilities to allow commanders and operators "to exercise appropriate levels of human judgment in the use of force" and that armed forces "employ systems with appropriate care and in accordance with the law of war, applicable treaties, weapon systems safety rules, and applicable ROE."[13]

Absent in the Directive is a definition or explanation of what the appropriate levels of human judgment are that should be exercisable, and exercised, by military commanders and operators before, during, and after the use of lethal force by autonomous machines in armed conflict.

I argue that, *legally*, prior to each individual attack using lethal force, the appropriate level of human judgment

is the attention necessary to ensure confidence that a selected autonomous weapon system complies with IHL and IHRL. Specifically, human interaction must be sufficient to satisfy the rules for targeting military objectives expressed in Articles 48-59 of API.[14] With present levels of artificial intelligence, autonomous weapon systems might meet this standard only in extremely limited circumstances.[15] As technology advances, however, autonomous weapons will enjoy greater capability to comply with legal obligations, and, in some situations, may out-perform humans in this regard.

In addition to the requirements of international law, political interests may compel the participation of humans in decisions concerning the use of force during warfare. Therefore, implementation of the Directive's "appropriate level of human judgment" standard requires a two-part analysis, including: 1) the identification of the international legal rules applicable to the use of force by an autonomous weapon and the capability of the weapon to act consistently with those rules, and 2) whether any political interests exist that might compel the exercise of particular levels of human judgment on the use of force by autonomous weapon systems.

The Legal and Operational Challenges of Autonomy. The advent of autonomous weapon systems creates options (and potentially the logistical, strategic, and legal need) for delegation and deferral of human responsibility and judgement to machines for the hard decisions that must be made to obtain military advantage, defeat the enemy, comply with IHL, and keep one's own forces alive.[16] Nevertheless, too much trust placed by humans in machines may endanger lives or yield other consequences. For example, autonomous systems containing the most sophisticated artificial intelligence—capable of applying knowledge and rules input by humans as well as self-appraisal similar to learning—may behave in unanticipated ways.[17] Furthermore, muddled lines of accountability may develop for particular acts or omissions, as soldiers and commanders effectively abandon their sense of responsibility for tasks, believing—or rationalizing—that a situation is "in the hands" of the computer.[18] As a result, soldiers and commanders may not fully understand the decisions they make or the actions they take.[19]

Appropriate levels of autonomy—and, conversely, human judgment—may vary depending upon the particular functions of a machine.[20] Every decision to modify a system's functions may require human judgment (at the design, testing, and/or operational stages) to determine the permissible level of autonomy accorded to the new task.[21] Thus, the person supervising the use and function of the weapons at a particular time and place must exercise sufficient judgment to ensure that the employment of the technology complies with IHL. Depending on the circumstances, such judgments may occur well in advance of a particular mission.[22]

The challenges of determining the appropriate levels of human supervision over autonomous weapon systems may grow exponentially as technology develops, in particular due to "the flood of collected data" that unmanned systems

may transmit to human overseers.[23] Commanders in the field—required to make split-second decisions where lives are at stake—simply cannot process all of the data now available to them in the virtual world. "There is information overload at every level of the military—from the General to the soldier on the ground."[24] Without careful consideration of available information, the danger exists that military officers will simply react to events, making snap decisions about the use of autonomous weapon systems without fully understanding their environment. In fact, they may not use their training and experience to make the best judgments in compliance with international law.

As Schmitt and Thurnher note, a "particularly critical issue to be considered is time."[25] During the planning and execution of lethal attacks by human operated weapon systems, decisions to continue or terminate the attack can sometimes be made up to the point of impact.[26] As the operational speed of new weapon systems technology increases, however, the time available for human reflection and judgment may be reduced to nothing, or nearly nothing.

One solution may lie in the development of even more technology, including powerful machines that have the capability to match or surpass human perceptual capabilities. For example, the U.S. government's Defense Advanced Research Projects Agency (DARPA) has an "XDATA" program dedicated to the development of "computational techniques and software tools for processing and analyzing the vast amount of mission-oriented information for Defense activities."[27] Ironically, perhaps, future evaluations about the appropriate levels of human judgment over autonomous weapon systems will depend on the use of developing technology. Human military judgment can become absent an effective system to collect, manage, analyze, and act on the vast amount of information available.[28]

Furthermore, the artificial intelligence used by autonomous weapons suffers from the "brittleness" problem; machines lack the ability to react in a reasonable time to dynamic and fluid environments.[29] Unfortunately, "war is the province of uncertainty."[30] To be effective, as both instruments of force and agents of international law, modern autonomous weapon systems must be able to react to changes in their environment that may not present themselves in stored computer algorithms or imagery.[31] The contemporary robotics technology available for lethal autonomous weapon systems is much too "brittle" to comply with IHL principles such as distinction and proportionality in a highly fluid and complex battle-space.[32] Therefore, the "brittleness" of present-day autonomous weapon systems usually requires the effective exercise of human judgment in the planning, execution, and evaluation of lethal, kinetic force.[33]

The Autonomy Continuum and the Place for Human Judgment. Contemporary writers often frame debates about levels of human attention to unmanned weapon systems in the relatively simple context of "man-in-the-loop," "man-on-the-loop," or fully autonomous weapon systems.[34] The difficulty with such labels is that they reveal little about the challenges

faced by persons and/or machines in understanding their environment, particularly during armed conflict. Thus, a so-called "man-on-the-loop" system tries to address the highly fluid and complex environments where the need for human judgement may be constantly shifting. It is important, therefore, to recognize that autonomy may progress along a continuum and different levels of human judgement may be appropriate for weapon systems that operate at different points along this gradation.[35]

At times, variations of autonomy become quite complex. For example, a single human supervisor may simultaneously be monitoring several computers or several computer networks. Varied computers or networks may offer the human supervisor conflicting suggestions for action, or different opportunities for human intervention prior to the computer taking action. Of course, each computer may be monitoring one or more autonomous weapon systems. These scenarios will only become more complex with the development of swarm technologies that will permit large numbers of robotic weapon systems to operate cooperatively and communicate rapidly amongst themselves. Furthermore, additional factors effect the exercise of human judgment such as the intent, orders, and influences expressed by commanders at different levels.

A distinguishing characteristic of future autonomous weapon systems, in addition to their independence, will be the speed with which these machines communicate information and execute decisions. This quality will generate opportunities for significant military advantages. It will also, however, further limit capacities for human command and control. Schmitt and Thurnher argue that "humans are never really 'out of the loop'" because "humans will decide when and where to deploy the [autonomous weapon] system and what parameters to embed within it."[36] Yet, when autonomous weapons react to events and use force at speeds that effectively prohibit human influence or intervention, soldiers, operators, and commanders are effectively "out of the loop."[37]

Accordingly, the development of increasingly powerful and fast autonomous weapon systems will force armed forces to choose between the potential for greater speed of action—"military advantage" in IHL terms—and the maintenance of human command and control.[38] Where the absence of human supervision during the employment of autonomous weapon systems would lead to violations of the principles of Distinction and Proportionality, or limitations on feasible measures of precaution in attack, the IHL rules of targeting would compel humans to engage with the weapon. On the other hand, as technology continues to evolve, autonomous weapon systems may fulfill obligations of IHL more successfully than humans.[39]

Nevertheless, Directive 3000.09's requirement that the operation of lethal autonomous weapon systems include "appropriate levels of human judgment" to fulfill not only the requirements of IHL but also ROEs suggests that the phrase "appropriate levels of human judgment" may extend beyond legal obligations. Thus, it is important to consider whether moral or political

reasons exist to compel the continued exercise of human judgment over autonomous weapons in circumstances where the machines can comply with IHL.

Political and Moral Values Compelling the Exercise of Human Judgment. There are a number of moral and political reasons why the exercise of human judgment over decisions by autonomous machines to use lethal force may be important, particularly during armed conflict. It is possible that lethal autonomous weapons may one day have the capability, using algorithms and machine learning, to apply the principles and rules of IHL to particular circumstances.[40] Nevertheless, the ability to strictly comply with the laws of war may not, by itself, make lethal autonomous weapon systems appropriate decision-makers during armed conflict. The law is not necessarily the final arbiter of the conduct of warfare.[41]

For example, in 2009, General Stanley McChrystal, then-commander of the International Security Assistance Force (ISAF) in Afghanistan, issued a tactical directive restricting the use of proportionate air support during combat against residential compounds and other locations where civilians may be present. McChrystal explained that this operational limitation was necessary "to avoid the trap of winning tactical victories but suffering strategic defeats by causing civilian casualties or excessive damage and thus alienating the people."[42] Similarly, a commander of forces in environments like Afghanistan might restrict the use of lawful yet lethal autonomous air and ground vehicles in combat operations due to strategic concerns whereby the presence or use of such deadly robots may alienate the local population. Or, in an additional effort to avoid unintended engagements, the force commander might require that lethal autonomous weapon systems be operated at all times in some version of the "man-in-the-loop" or "man-on-the-loop" modes. The possibility exists that, in particular circumstances, autonomous weapon systems may have the capacity to fulfill legal requirements and accommodate specific ROEs. Nevertheless, the "appropriate levels of human judgment" for compliance with ROEs may be greater than those required for compliance with international law.

Arguably, in addition to strategic or political goals, moral values grounded in the principle of humanity might also compel the exercise of higher levels of human judgment over lethal autonomous machines. First, the violence of war is brutal and usually requires the suppression of much in our nature that is human and humane. Nevertheless, during warfare the suppression of humanity and identity by human combatants is occasionally offset by (often irrational) expressions of compassion and acts of courage.[43] This ability to empathize with human beings, whether soldier or civilian, would be absent from robotic weapons.[44] Removing these human elements from military decision-making processes—and replacing them with predictable but soulless mechanical intelligence—risks losing the profoundest manifestations of humanity.[45]

Second, there will be a significant cost to society if we lose the identity

of "warriors." If we make obsolete the social and professional group whose identity includes an understanding that "war is inevitably tragic, inefficient, and uncertain" the veil of technology will transform war into an experience with few lessons or insights.[46] Yet, "insight is what saves us from ourselves."[47] The human cost of the developments and employment of autonomous weapon systems is a reduced understanding of the destructive violence of armed conflict.

Similarly, one might argue that delegating decision-making power from humans to machines inevitably distances human beings from the consequences of their actions.[48] This will be a dangerous phenomenon that will undermine the value of human conscience, in particular the sense of responsibility for fellow human beings that is important in any healthy society.

None of these moral arguments appears to be particularly compelling, however, for the simple fact that human beings often do not exercise these responsibilities in ways that satisfy moral values. Reflecting upon the Holocaust, Primo Levi concluded, "the true crime, the collective general crime of almost all Germans of that time was that of lacking the courage to speak."[49] Hannah Arendt observed that perfectly normal persons may be completely incapable of distinguishing right from wrong, or skilled in coloring wrongful behavior in benign terms.[50] According to Ervin Staub, "evil that arises out of ordinary thinking and is committed by ordinary people is the norm, not the exception."[51] Thus, both moral duty and legal obligation during armed conflict imply forms of responsibility based on often-flawed human perceptions.

Autonomous weapon systems may indeed alter moral perceptions of situations during armed conflict in ways that are difficult to foresee today. For example, victims and witnesses of traumatic circumstances, such as soldiers, often "try, usually in vain and at great expense of energy, to banish what has happened to them from their minds."[52] If robot soldiers may be designed to exclude "the anguish of memory," perhaps the suffering caused by war—to combatants and civilians alike—might be reduced.[53] In 2011, Jacob Kellenberger, President of the International Committee of the Red Cross, observed that a "robot could be programmed to behave more ethically and far more cautiously on the battlefield than a human being."[54]

A deeper examination of this argument, however, uncovers its weakness. The unpleasant memories of war can serve as an important future brake to the escalations of violence that result in armed conflict. Thus, one important element possibly underlying the "human judgment" requirement in DOD Directive 3000.09 is that sound human judgment often evolves from life's difficult and painful experiences.

It takes time, however, for humans to experience, process, and adjust to the challenging experiences of their lives. None of the moral arguments mentioned above for maintaining greater levels of human judgment override the most basic criteria of military necessity: the simple need of soldiers and armies to survive. Therefore, moral values alone are unlikely to support a requirement for higher levels of human supervision during the operation of

autonomous weapon systems.

Conclusions. Although less than perfect, the DOD Policy Directive represents one of the few national attempts to articulate autonomous weapon policy. Any future policy directives or guidelines on the employment of autonomous weapons systems should explain that, to comply with international legal obligations or political interests, the intervention of human judgement may be required at three distinct stages of military operations: 1) at the planning stage of the mission or attack when a human must choose which weapon system to employ; 2) following the choice of an autonomous weapon system, a decision must be made as to the level of human attention to assign to the system during the task prior to an attack; and 3) specific inputs of human judgement immediately before, during, and after the attack. Human control over the autonomous weapon system should remain until the human supervisor is confident that, at each of these three stages, the weapon system complies, and will continue to comply, with international law and applicable rules of engagement.

More complex tasks and more limited autonomous technology will signal a demand for greater appropriate levels of human judgment and communication during the mission. The opportunity for humans to apply inductive and creative reasoning may produce more effective weapon systems.[55]

The challenge of maintaining appropriate levels of human judgment and adherence to international law may relate less to the level of autonomy of a lethal autonomous weapon system and more to the speed with which humans and future technology can absorb, process, and transmit information and react to events. Nevertheless, the development of fast and lethal autonomous weapon systems should not supersede the legal and political value of human judgment in the important decisions determining the conduct of armed conflict.

NOTES

1 IHL is also known as the law of armed conflict or the law of war. ICRC, "What Is International Humanitarian Law?" Internet, http://www.icrc.org/eng/resources/documents/legal-fact-sheet/humanitarian-law-factsheet.htm.

2 See Alejandro Valencia, *La Humanización de la Guerra: Derecho Internaciónal Humanitario y el Conflicto Armado en Colombia*, (Bogota: Ediciónes Unidanes, 1991), 202.

3 Nevertheless, the increasing use of technology by humans has driven the development of the laws of war. Theodore Meron, *Bloody Constraint: War and Chivalry in Shakespeare* (Oxford: Oxford University Press, 1998), 12.

4 With respect to weapons that are generally defensive in nature, modern militaries already cede a great deal of independence for the use of lethal force. Many navies, for example, operate the Phalanx Close in Weapons System on their ships against air warfare threats such as planes and missiles. The U.S. Navy describes Phalanx as "the only deployed close-in weapon system capable of autonomously performing its own search, detect, evaluation, track, engage and kill assessment functions." Use of systems such as Phalanx to defend ships at sea implies very limited risks for civilians. See, "Phalanx Close-In Weapons System," Internet, http://usmilitary.about.com/library/milinfo/navyfacts/blphalanx.htm.

5 Ashton B. Carter, "Autonomy in Weapons Systems," *Department of Defense Directive, United States of America, Number 3000.09* (2012).

6 This definition "includes human-supervised autonomous weapon systems that are designed to allow human operators to override operation of the weapon system, but can select and engage targets without further human input after activation."

7 Ashton B. Carter, "Autonomy in Weapons Systems," *Department of Defense Directive, United States of America, Number 3000.09* (2012), 14.

8 Ibid, Enclosure 4, "Responsibilities," parts 1 - 10.

9 Ibid, 2.

10 Ibid.

11 Ibid.

12 Ibid.

13 Ibid, 1.

14 "API" refers to the first 1977 Additional Protocol to the Geneva Conventions of 1949. In general terms, the targeting rules of IHL address the concepts of Distinction, Proportionality, and Precautions in Attack.

15 Bill Boothby, "How Far Will the Law Allow Unmanned Targeting to Go?" in *International Humanitarian Law and the Changing Technology of War*, Dan Saxon, ed., (Leiden: Martinus Nijhoff/Brill, 2013), 57-59.

16 Armin Krishnan, *Killer Robots: Legality and Ethicality of Autonomous Weapons* (Surrey: Ashgate Publishing, 2009), 86.

17 Gianfranco Visentin, Interview, Automation and Robotics Section, European Space Agency, Noordwijk, The Netherlands, 4 November 2013.

18 Sheridan calls this dynamic the "phenomenon of misplaced trust." See, T.B. Sheridan, et. al., "Adapting Automation to Man, Culture and Society," *Automatica* 19 (1983): 610.

19 I am grateful to William Boothby for this insight.

20 Martin Azkarate, Interview, Automation and Robotics Section, European Space Agency, Noordwijk, The Netherlands, 15 October 2013.

21 William Boothby, Interview, Brookwood, U.K., 3 October 2013.

22 Ibid.

23 "The U.K. Approach to Unmanned Aircraft Systems," *Joint Doctrine Note 2/11*, 30 March 2011.

24 Thom Shanker and Matt Richtel, "In New Military, Data Overload Can Be Deadly" *The New York Times*, 17 January 2011.

25 Michael N. Schmitt and Jeffrey S. Thurnher, "Out of the Loop, Autonomous Weapons and the Law of Armed Conflict," *Harvard National Security Journal* 4 (2013): 268.

26 Ibid.

27 DARPA Information Innovation Office, "XDATA," Internet, http://www.darpa.mil/Our_Work/I2O/Programs/XDATA.aspx.

28 Darren Stewart, "Maximizing Compliance with IHL and the Utility of Data in an Age of Unlimited Information: Operational Issues," in *International Humanitarian Law and the Changing Technology of War*, Dan Saxon, ed., (Leiden: Martinus Nijhoff/Brill, 2013), 172.

29 Missy L. Cummings, "Human Supervisory Control of Swarming Networks," Internet, http://web.mit.edu/aeroastro/labs/halab/papers/cummingsswarm.pdf, 3.

30 Carl von Clausewitz, *On War* (Harmondsworth: Penguin Books Ltd, 1968),140.

31 Damian A. Green, *The Future of Autonomous Ground Logistics: Convoys in the Department of Defense* (Fort Leavenworth, KS: School of Advanced Military Studies, United States Army Command and General Staff College, 2011), 9.

32 Martin Azkarate, Interview, 15 October 2013.

33 See Bill Boothby, "How Far Will the Law Allow Unmanned Targeting to Go?" in *International Humanitarian Law and the Changing Technology of War*, Dan Saxon, ed., (Leiden: Martinus Nijhoff/Brill, 2013), 63.

34 David Akerson, "The Illegality of Offensive Lethal Autonomy," in *International Humanitarian Law and the Changing Technology of War*, Dan Saxon, ed., (Leiden: Martinus Nijhoff/Brill, 2013), 71-72.

35 Gianfranco Visentin, Interview, 4 November 2013.

36 Michael N. Schmitt and Jeffrey S. Thurnher, "Autonomous Weapon Systems and the LOAC," *Harvard National Security Journal* 4 (2013): 280.

37 Presently military "communication occurs at the speed of light" and recognizes that autonomous weapons systems permit "the speed of future deci-

sion cycles outpacing the human mind." M. Shane Riza, *Killing Without Heart: Limits on Robotic Warfare in an Age of Persistent Conflict* (Washington, D.C.: Potomac Books, 2013), 41.

38 For example, the full utility of multiple, cooperating autonomous aerial weapons systems "will only be realized if they can operate autonomously: able to fly, adapt, communicate, negotiate, and carry out missions *with no human in the loop*." Robert J. Bamberger Jr., et. al, "Flight Demonstrations of Unmanned Aerial Vehicle Swarming Concepts," *Johns Hopkins APL Technical Digest* 27 (2006): 53.

39 *International Humanitarian Law and the Changing Technology of War*, Dan Saxon, ed., (Leiden: Martinus Nijhoff/ Brill, 2013), 349.

40 *Joint Doctrine Note 2/11*, para. 507.

41 Michael N. Schmitt, "Targeting and International Humanitarian Law in Afghanistan," Israel Yearbook on Human Rights 39 (2009): 329.

42 HQ ISAF, "Tactical Directive," Internet, http://www.nato.int/isaf/docu/official_texts/Tactical_Directive_090706.pdf.

43 See Robert Chantler, "Translator's Introduction," in Vasily Grossman, *Life and Fate* (London: Flamingo, 1985),12.

44 Human Rights Clinic of Harvard Law School, *Losing Humanity, the Case Against Killer Robots* (2012), 38.

45 Delegating the process of using lethal force "dehumanizes armed conflict even further and precludes a moment of deliberation in those cases where it may be feasible." See "Lethal Autonomous Robotics and the Protection of Life," in *Report to the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions*, A/ HRC/23/47, 9 April 2013.

46 Greg Jaffe, "Former Defense Secretary Gates Warns Against Lure of Drone Warfare," *The Washington Post*, 24 October 2013.

47 Robert Stone, "Introduction" in Michael Herr, *Dispatches* (New York: Alfred A. Knopf, 1977), xiii.

48 Nick Bostrom and Anders Sandberg, "The Future of Identity," (London: UK Government Office for Science, 2011), 17.

49 Primo Levi, *The Drowned and the Saved* (Italy: Einaudi, 1986),182.

50 Hannah Arendt, *Eichmann in Jerusalem: A Report on the Banality of Evil* (New York: Penguin Books, 1977), 26, 69.

51 Christopher R. Browning, *Ordinary Men: Reserve Police Battalion 101 and the Final Solution in Poland* (New York: Harper Collins Publishers, 1998), 167.

52 W. G. Sebald, *On the Natural History of Destruction* (London: Hamish Hamilton, 2003), 153.

53 Ibid. The exclusion of memory, however, may be a double-edged sword as autonomous weapon systems with learning capabilities will require some ability to record and evaluate the changing circumstances around them.

54 *International Humanitarian Law and the Changing Technology of War*, Dan Saxon, ed., (Leiden: Martinus Nijhoff/ Brill, 2013), 27.

55 Missy Cummings, "Man v. Machine or Man Plus Machine?" (forthcoming).