WHEN MACHINES KILL: CRIMINAL RESPONSIBILITY FOR INTERNATIONAL CRIMES COMMITTED BY LETHAL AUTONOMOUS ROBOTS

Oren Gross[1]

The prospects of what may be regarded as practically risk-free wars either among clone armies or when one side controls Lethal Autonomous Robots (LAR) raise significant legal, ethical and policy questions about the nature of warfare in the future as well as the appropriate regulation of future wars.

The first question that needs to be tackled is one of definitions. Much of the literature (and there is not much to start with) refers to 'autonomous weapons systems' (AWS). This terminology raises such questions as to what do we mean by 'autonomous' as well as the critical question of whether such systems are to be regarded as 'weapons'—and thus subject to rules, limitations and restrictions pertaining to weapons and the ways in which those may be used—or as something else (and if so, what). This has obvious implications for issues of criminal liability as we consider where on the spectrum between the knife and the action of a truly independent (morally and legally) agent we place

[1] Irving Younger Professor of Law and Director, Institute for International Legal & Security Studies, University of Minnesota Law School.

the LAR. The answer to this initial question depends, to a large extent, on the nature of the system that we examine. While land and sea mines may be regarded as examples of AWS and even LAR I would suggest that they are very close to the knife end of the spectrum. On the other hand we are coming ever closer to the point in which truly independent LAR will operate in the battle space. When that is the case questions of criminal liability for crimes committed by such systems are bound to emerge. Not surprisingly international law does not offer any clear and agreed upon definition of LAR (nor, for that matter, does any domestic legal system that I am aware of).

For the purposes of my paper I regard as an LAR any system that is capable of carrying out lethal attacks without human intervention in the process, i.e., with no human operator in-the-loop. Thus, weapons systems that have attracted much attention and have been subject to recent public and academic debates such as unmanned aerial vehicles (also known as drones), guided missiles and similar tele-operated weapons systems are not LAR. While tele-operation may mean that the system in question is not a true LAR, questions still rise with respect to other mechanisms of possible human control over the actions of the system (e.g., abort mission commands or kill-switches). Again, degrees of control will be relevant to determination of criminal liability both under international law and domestic law. For my purposes, true LAR will

be capable of determining for itself, that is without human input (I leave aside the questions of if and when we may cross the threshold of machine self-awareness which may necessitate conceptualizing such systems in human terms rather than as inanimate objects), the means and methods by which to carry out a given mission. It would be capable of identifying targets and carrying out military operations including, significantly, lethal attacks. It will have the decision-making capacity to identify a target, determine whether (or not) to use lethal force against it, decide what type of weapons (from a selection of payloads it carries) to use should it decide to attack etc.

Once LAR can appear in battle-space (the issue is not if but rather when) questions of their capacity to comply with the laws of war would emerge. And so will questions of criminal liability for their operations. What possible bases of criminal responsibility may be applied when LAR commit international crimes? Will criminal defenses such as self-defense be available in the context of LAR operations?

There exists an inherent tension between the degree of autonomy that LAR enjoy and issues of accountability for their actions. This tension is exacerbated further in light of what may be seen as the trend not only towards the distancing (spatially and temporally) of human combatants from the battle space (e.g., swords replacing knives, guns replacing swords, artillery and indirect

line of fire projectiles, guided missiles, mines, drones) but, indeed, the future removal of humans from the scenes of war. While, again, this offers the double-edged sword of a risk-free warfare (e.g., reducing human suffering—both for soldiers on the field and for civilians because "Machines…don't have hate, vindictiveness, cruelty, or psychosis. Machines don't rape innocent women. Machines don't abuse prisoners. Machines don't massacre civilians"[2] [Sharkey]—but also lowering the costs of going to war [both monetary costs and cost in terms of potential human suffering] and lowering incentives to terminate armed conflicts), it also bespeaks of a paradigm-shift towards wars that are, literally, dehumanized or, perhaps, "post-humanized."

Whether criminal law can deal adequately with such challenges (and if so how) remains to be seen. The ways in which international humanitarian law (with its continuous balancing of military necessity and humanitarian concerns) can regulate, govern and deal with these new technological advancements—with the potential introduction not merely of new means and methods of

[2] I note that far from being factual, these statements are controversial. **First**, they assume neutral, value-free, and comprehensive programming which may not be the case. Thus, it is quite possible that machines will be programmed to massacre civilians. **Second**, if machines lack hate and other negative emotions so too do they lack empathy and positive emotions that may be critical on the battlefield (e.g., refusal to commit a war crime). Recent studies argue that far from contributing to irrational decision-making, emotions are part of our rational self and play a significant role in enabling real-time practical thinking [e.g., Damasio's experiments involving human beings with brain damage]. **Finally**, I am greatly troubled by the flippant and careless statement that "machines don't rape innocent women" as if raping women who are somehow not innocent may be justified or excused.

warfare but of entirely new objects and (potential) subjects of warfare—also remains to be seen.

Conceptually we should inquire to what extent is *human* responsibility and accountability a necessary condition for the systemic compliance with the laws and customs of war. It seems to be clearly the case today that such responsibility is necessary. But would the introduction of LAR to the battle-space change any of that? Should it? If it does then it creates strong incentives to utilize such systems in order to distance humans from responsibility. On the other hand, the greater the degree of autonomy of such systems—the more they come to resemble human actors—the more critical becomes the inquiry about the scope and confines of human responsibility (again, if we are in a real futuristic mode as are those who, for example, subscribe to the singularity theory, we should contemplate the possibility that the time may come when imposing criminal responsibility on LAR may, in fact, be sufficient since the thinking machine will, for practical purposes, be "human"). Whichever way we come out on these questions, we should consider the question of whether—assuming that no human being could be held criminally responsible for crimes committed by LAR—it would be ethical to use such systems in the future battle space and whether such use ought to be (somehow) outlawed and made unlawful.

It seems that another necessary inquiry concerns what it is that makes us instinctively recoil against the specter of autonomous machines, in general, and in the battle space in particular. In other words, what human attributes we deem as indispensible for the conduct of war. Prime candidates seem to be discretion, judgment, accountability (both moral and legal), and (for better or worse) feelings. Leaving aside the issue of accountability, it remains an open (and clearly controversial) question to what extent robots in general and LAR in particular may, in the future, be able to out-perform humans in their capacity to ethically (in the sense of rule-based) use lethal force on the battlefield. In an increasingly complex battle-space in which paradoxically time and space contract (paradoxically in light of the trend to move humans further away from the battlefield) and in which the need for split-second information gathering and processing (and decision-making) becomes increasingly more relevant it is not entirely clear that humans would perform better (again in the sense of rule based conduct) than LAR (e.g., the downing of Iran Air flight 655 by the U.S.S. Vincennes).

Before exploring these questions further I put forward four assumptions that inform my discussion below: **(1)** The use of LAR is not done with the aim of shielding humans (whom ever they may be) from criminal responsibility; **(2)** The use of LAR is not done with the aim of committing a war crime, crime of

aggression or a crime against humanity; **(3)** It is feasible to program LAR in accordance with the laws and customs of war, both Hague and Geneva laws (a question that needs further exploration here is what would be required to make such programming "feasible"). In this context one should particularly focus on the principles of discrimination (between combatants and non-combatants, military and non-military targets) and proportionality. Clearly, deploying LAR that cannot discriminate combatants from civilians would, in and of itself, violate the laws of war (much like the use of a weapon that is not capable of discriminate targeting). Whether or not we could program LAR to so discriminate remains at this point an open question. Similar issues arise in the context of the principle of proportionality. Finally, **(4)** the discussion below pertains to a known extent to a transitional period (whose duration remains, at this point, unknown) stretching between the present and that point in the future in which "the Singularity" will occur. In the sciences, singularities signify situations that lead to a radical change in understanding, concepts and conceptions, and rules. Normal, ordinary rules and principles do not apply to the Singularity—such as a black hole—and beyond it. A tectonic paradigm shift results in the change of much of what we know. As Singer puts it aptly, "[t]he key is that someone living in a time before a paradigm shift would be unable to understand the world that follows." Adherents of the Singularity thesis in our

context believe that at a certain point in the future (estimates as to how near that point might be vary greatly) artificial intelligence systems will become akin to a new species. It may match the capacities of the human brain and even surpass them and, significantly, may become self-aware or HAL-like. When (and according to others, if) that point is reached the question of criminal liability and the significance of attaching responsibility to human beings may shift again as robots and AI will become human-like.

The remainder of this brief paper looks, therefor, at the potential models for establishing criminal responsibility for international crimes committed by LAR during that transitional period.

Holding the LAR directly responsible: This seems to raise several distinct questions. **First**, what would be required in order to find LAR criminally liable? Significantly, in this context we need to consider how LAR are to be conceptualized and understood: should they be regarded as (sophisticated to be sure) weapons to which the rules of Art. 8(2)(b)(xx) of the Rome Statute (as well as the body of laws that deal with the permissibility of weaponry) apply or rather should they be thought of as agents capable of being directly criminally liable? At present we may say that machines have no capacity to "want" to kill civilians as they have no independent "desires" [Singer: 388]. Thus machines seem incapable of being responsible for war crimes. Yet, I

would suggest that this may not hold with respect to future LAR who may have their own sets of goals—machine-desires—and carry out their actions in accordance with, and in furtherance of, those goals. It should be noted that if LAR are absolved of responsibility for the commission of war crimes (and assuming that they distance humans from such responsibility) there will be strong incentives for states to use such LAR in the battle-space (in which case the first two assumptions laid out above will not hold).

Second, what sort of criminal sanctions would be available in case the LAR is found legally responsible? How do theories of punishment fit the case of the LAR (e.g., is there any point to talk about deterrence Or rehabilitation?)?

Third, could LAR invoke defenses against criminal responsibility? E.g., software malfunction (due to problems in the design thereof or to the intervention of extraneous elements such as computer viruses) as, perhaps, equivalent to various situations of incapacitation; Self-defense when the "self" is not a human being; Duress; Mistake of fact.

In this context it is interesting (and provocative) to note that some writers have equated AWS (not drawing a clear distinction between those and true LAR) with child-soldiers.

Finally, according to Article 25(1) of the Rome Statute, the International Criminal Court has jurisdiction only over "natural persons." This provision, which is almost deemed axiomatic is clearly designed to distinguish the jurisdiction of the ICC from that of international tribunals such as the ICJ (jurisdiction over states). Yet, in the context of LAR that jurisdictional clause raises obvious challenges (assuming, of course, that one could speak of criminal responsibility of LAR).

Voluntary intoxication: Could the use of LAR be analogized to self-induced intoxication that would not relieve the actor from criminal responsibility? Could it be argued that the mere use of LAR is analogous, in other words, to putting oneself in a situation in which the actor has no control over his or her actions (after all that is not that the point behind the autonomy component of the LAR)? If that is the case we ought not to allow the military (or other governmental agencies for that matter) to be able to distance itself from the crimes committed by the LAR. However, such a construction seems highly problematic as well. **First**, if my assumptions hold then the mere deployment of LAR does not amount to a wrongful act. This is not a scenario in which the military uses LAR precisely in order to "lose control" and detach its officers and soldiers from responsibility for the actions of the LAR (and,

after all, we assume the feasibility of programming LAR in accordance with the laws and rules of war).

Second, if the third assumption holds it is certainly arguable that deploying LAR demonstrates greater commitment to the laws of war than deploying human soldiers. After all, if programmed 'correctly' the LAR is far less likely to act in violation of these laws than any human being (e.g., it is capable of assessing the relevant facts with greater speed and accuracy and is unlikely to be moved by negative emotions such as fear, panic or hatred). LAR could, arguably, "preserve the legitimacy of the cause because the use of force is constrained by a rigid set of heuristics preprogrammed to comply with [the rules of engagement and the laws of armed conflict]." [Guetlein (2005): 11].

Perpetration-by-means [of another] (PMA): refers to offenses committed by and through an innocent agent who (or which) is not criminally responsible (e.g., a minor or a mentally incompetent person). The acting agent is deemed analogous to an instrument that is utilized by the real perpetrator, the principal, who is held responsible for the actions of the agent.

In our context PMA raises several questions. If we follow the PMA model, who should be held criminally liable as the 'principal'? Do we hold the programmer of the LAR liable? Or perhaps the military commander on the

scene (how high up do we go?) under whose "command" the particular LAR operates? Or the "operators" for the particular mission in which the alleged crimes had been committed?

Could PMA be applied to the case of LAR? **First**, the PMA model requires a degree of control over the innocent agent that may well be missing in the case of a LAR. Again, consider the main characteristic of LAR, i.e., its autonomous decision-making process. That autonomy means that its decisions are not necessarily predictable (unless we believe that preset rules capture the full panoply of possible scenarios and questions that may arise) and that in arriving at them LAR operates without human intervention or influence. LAR is capable in "making up its mind" and deciding how to pursue "the mission." After all "if a machine is expected to be infallible, it cannot also be intelligent" [Penrose (1989): 124]. If all eventualities and all answers to all situations are keyed in (assuming this is even possible) then there is no real point in speaking of an "intelligent" machine as opposed to one that merely parrots what its preprogrammed to do. Thus notions of intelligence and, I would argue, autonomy, suggest "heuristic" rather than "full-proof" programming which means, in turn, that the machine will take certain short-cuts in its decision-making process in ways that may leave it open to error, however low the probability [Dennett (1997): 363].

While existing AWS (I use this term here consciously) are rather simple affairs of the "fire and forget" variety, the LAR of the future will be able to make independent decisions (within the framework of pre-specified rules) on a broad range of issues. Significantly, these systems are also likely to be self-correcting in the sense that they will learn from experience (their own and other systems') and as such are going to move even further from their initial programming.

Second, what could the 'principal' be charged with? It would seem that he or she should have "the kind of culpability that is sufficient for the commission of the offense." [Robinson (1983): 734]. What would be the mens rea of the 'principal' in the context of operations by LAR? (e.g., in the case of the programmer, operations that may take place many months or years after the LAR was commissioned); What could the 'principal' be charged with?

Third, is LAR really an "innocent" and is it really an "agent"? Could a LAR be held criminally responsible (whatever that may mean in the context of a robot) for its own actions? Is it like a stone or a dog [Dressler (1989): 1359] or a child or a legally incompetent adult, or, perhaps, more like an independent moral agent? On the one hand, the traditional view had been that PMA requires acting through an innocent agent or "causing crime by an innocent" [E.g., MPC §2.06(2)(a)]. On the other hand, it seems that some jurisdictions are

willing to accept a construction of a "perpetrator behind a perpetrator" [Fletcher, Basic Concepts of Criminal Law: 198-99].

Article 25(3)(a) of the Rome Statute incorporates the possibility of criminal liability for crime committed "through another person, *regardless of whether that other person is criminally responsible*." This provision was discussed in the cases of Lubanga, and Katanga and Chui as well as invoked by the ICC Prosecutor in the application for the issuance of an arrest warrant against Omar Al Bashir.

The possibility that the concept of indirect perpetration could become, in general, a key mode of liability in international criminal law [Jessberger & Geneuss (2008)] returns our attention to the question of the degree of control over the agent. Is the case of LAR analogous to the case of the East German border guards and the members of the GDR's Security Council who gave the guards the 'authorization to shoot'? After all, the LAR may not get a specific 'authorization to shoot' but rather a mission couched in more general terms and the control over the guards may not be the same as the 'control' of the LAR.

Furthermore, the autonomy of the LAR implies that in many cases it will determine on the means and methods of carrying out its mission independently of any human input and influence. In other words, the decision to act will be *its*

rather than the programmer's or the relevant military authorities. The LAR will decide to fire on a target that it would identify as a legitimate target and would factor in other relevant factors (e.g., collateral damage) based data that it would gather (or receive from other mechanisms of information gathering, including other LAR and warbots) and process. The commission of a particular act will be its own as an independent perpetrator rather than as an agent.

Fourth, and related to the point made above, what is the scope of application of PMA compared with doctrines such as criminal responsibility of commanders or other superiors? To what extent could PMA be applied to a hierarchical organization such as the military or an organized criminal group in order to "go after" the military commanders or arch criminals who operate behind the scenes and who may not have a direct, specific link to the commission of concrete offenses but rather set the "guidelines" for operations in general [Gur Arye (2011)]? In other words, could domination and control that are arrived at through effective organizational hierarchies substitute for the traditional concepts [Roxin; Jessberger & Geneuss]? Even if one subscribes to this extension of the doctrine, it seems that the rationale for the expansion is still found in the ability of the 'principal' to impose her will on the actor (whether due to the "innocence" of the latter or due to the organizational structures within which the two operate). Is that the case with LAR?

***Multi-offender crimes**:* Consider the relationship between PMA and alternative models of criminal responsibility for multi-offender crimes. **First**, *co-perpetration* (Article 25(3)(a) of the ICC Statute speaks of commission of a crime "jointly with another"). However, the case of LAR does not seem amenable to a horizontal attribution of criminal responsibility and it is unlikely to establish an agreement (explicit or implicit) between humans and LAR to commit the crime. **Second**, *ordering, soliciting or inducing a crime* (Article 25(3)(b)) seems inadequate inasmuch as the autonomous nature of LAR would mean (see the basic assumptions) that the crime actually committed or attempted is not one that had previously been ordered, solicited or induced. At the same time, this model of responsibility may be useful inasmuch as it does not require humans to be in control of the LAR when it commits the crime. **Third**, *aiding, abetting or assisting* (Article 25(3)(c)) would mean turning LAR into the principal with criminal responsibility for humans being secondary and derivative. This takes us back to the conceptual questions about the nature of LAR and their own criminal responsibility. **Fourth**, Article 25(3)(d) refers to contribution to a *crime perpetrated by a group of persons with a common purpose.* As this provision is the "little cousin" of the doctrine of joint criminal enterprise [Weigend (2008): 478] that has received much criticism in cases such as Lubanga, it seems an unlikely candidate for the purpose of imposing criminal liability on the humans that

may be deemed involved in the operation of the LAR. **Fifth**, *superior responsibility* (Article 28) is rejected by some scholars on the basis that it establishes responsibility for omission [Jessberger and Geneuss (2008): 865]. Substantively, it is not clear that the elements for superior responsibility under Art. 28 exist (again, under my three assumptions). Thus, it is not at all clear that we could speak of the failure by a military commander to "take all necessary and reasonable measures within his or her power to prevent or repress [the] commission [of the crime]…" in light of the argument (as noted above) that deploying LAR that are more rule bound than human soldiers. Finally (and obviously), this venue of attributing criminal responsibility would not be applicable to the programmers.

Causation: could we find a legally significant causal link between the programmer or the relevant military authorities and the harmful result due to the actions of the LAR? If they actually foresaw or could have foreseen that the LAR would violate the laws of war and commit a crime the answer seems straight forward enough. Once again we return to the question of how feasible it is to program and operate the LAR in accordance with the laws of war (with their inherent ambiguity at times).

Furthermore, would we need to show that the human agents could have foreseen the commission of the *specific* offense committed or would it be

sufficient to show that the commission of offenses was foreseeable? If it is the former it would be extremely difficult if not outright impossible to establish their criminal responsibility.

If LAR does, in fact, act in an independent, autonomous fashion, does this not disconnect the causal link between human agents and the crimes committed by the LAR? What about a situation in which the LAR decides to override a human commander orders and act in violation of the laws of war, e.g., killing surrendering enemy soldiers on the basis of calculating the mission costs [Sparrow (2007)]?