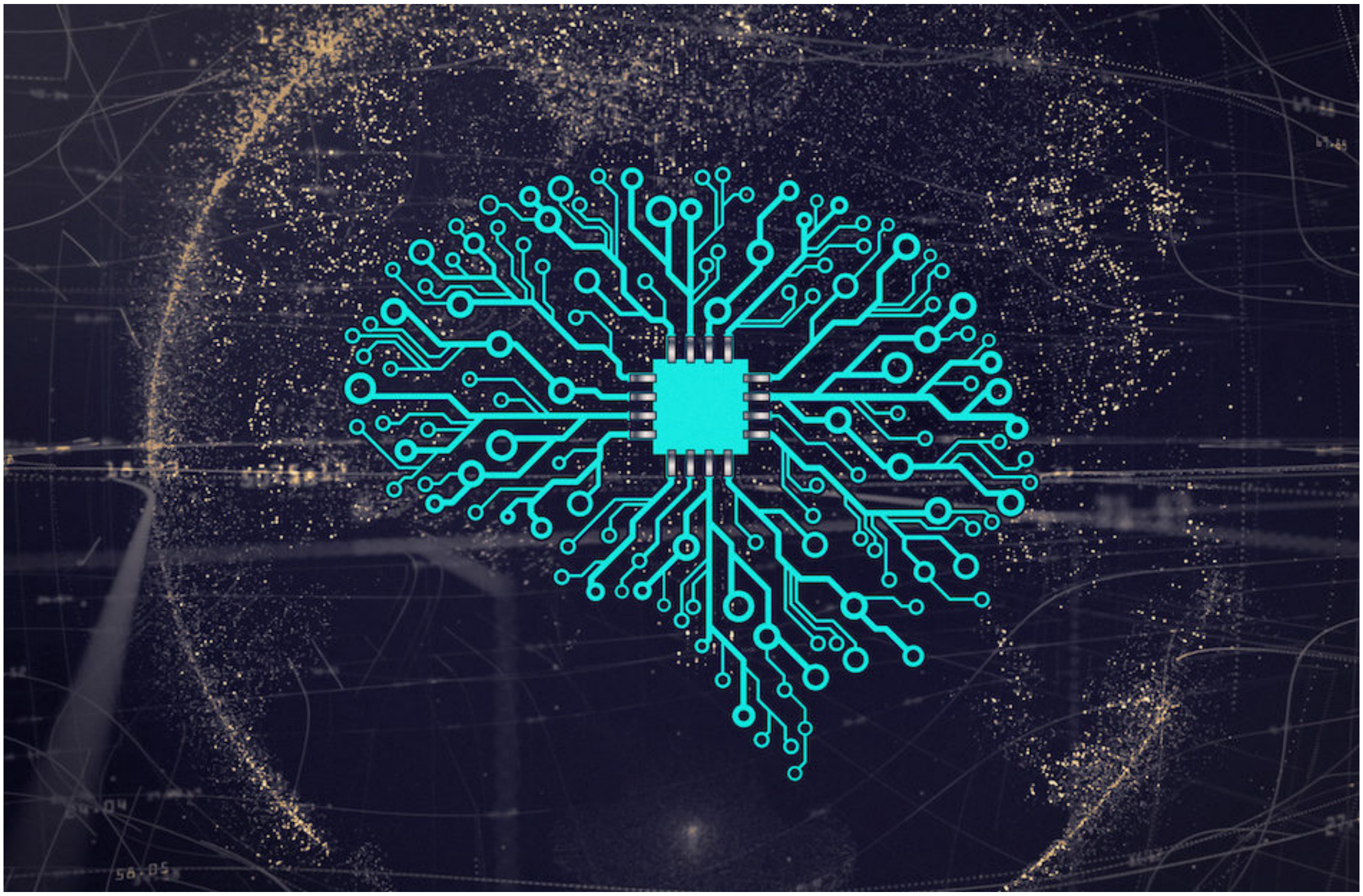
Legal reviews of weapons, means and methods of warfare involving artificial intelligence: 16 elements to consider

Legal reviews of weapons, means and methods of warfare involving artificial intelligence: 16 elements to consider

March 21, 2019, Artificial Intelligence and Armed Conflict / Conduct of Hostilities / Law and Conflict / New Technologies / Weapons

Dustin A. Lewis



What are some of the chief concerns in contemporary debates around legal reviews of weapons, means or methods of warfare involving techniques or tools related to artificial intelligence (AI)? One session of the December 2018 *workshop* on AI at the frontiers of international law concerning armed conflict focused on this topic. In this post, I outline a few key threshold considerations and briefly enumerate 16 elements tha States might consider as part of their legal reviews involving AI-related techniques or tools.

It is imperative, in general, for States to adopt robust verification, testing and monitoring regimes as part of the process to determine and impose limitations and—as warranted—prohibitions in respect of an employment of weapons, means or methods of warfare. Where AI-related techniques or tools are—or might be—involved, the design and implementation of legal review regimes might pose particular kinds and degrees of challenges as well as opportunities. With respect to challenges, for example, in a forthcoming blog post Netta Goussac will highlight several legal and other concerns that might arise in respect of reviews of weapons involving AI, not least the potential to introduce uncertainty and corresponding issues regarding (un)predictably and (un)reliability. Furthermore, today it seems, from my perspective, that sufficient trust among States in this area seems to be lacking, at least among certain States with advanced technological capabilities. Against that background, robust legal reviews may not only contribute to legal compliance, but may also help foster normative stability and augment trust among States.

What do I mean by AI-related techniques and tools?

But first, a word on what I mean by AI-related techniques or tools. My starting point is that there is no generally recognized definition of AI. That said, it might be of value to focus on techniques or tools derived from, or otherwise related to, AI science broadly conceived. My understanding —drawn from the work of such scholars as Barbara J. Grosz—is that AI science pertains in part t the development of computationally based understandings of intelligent behavior, typically through two interrelated steps. One of those steps concerns the determination of cognitive structures and processes and the corresponding design of ways to represent and reason effectively. The other step relates to the development of theories, models, data, equations, algorithms and/or systems that embody that understanding.

So defined, AI systems are typically conceived as incorporating techniques—and leading to the development of tools—that enable systems to 'reason' more or less 'intelligently' and to 'act' more or less 'autonomously'. The systems might do so by, for example, interpreting natural languages and visual scenes; 'learning' (or, perhaps more commonly, training); drawing inferences; and making 'decisions' and taking action on those 'decisions'. The techniques and tools might be rooted in one or more of the following methods: those rooted in *logical reasoning* broadly conceived, which are sometimes also referred to as 'symbolic AI' (as a form of model-based methods); those rooted in *probability* (also as a form of model-based methods); and/or those rooted in *statistical reasoning and data* (as a form of data-dependent or data-driven methods).

Existing and purportedly new or emerging primary norms

By way of reminder, under international humanitarian law/law of armed conflict (IHL/LOAC), *Article 36* of Additional Protocol I of 1977 provides that

> [i]n the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party.

What is the legal nature of these reviews? A determination of lawfulness, or lack thereof, by a State in respect of those treaty provisions is not—at least according to the Rapporteur of those provisions' drafting Committee (see *O.R. XV*, p 269, CDDH/215/Rev.1, para 30)—binding internationally. If we assume that that position is accurate, it would seem that the same contention might hold for the customary law counterparts, if any, of those treaty provisions. Instead, these legal review provisions—whether of a treaty or customary nature—might be seen as boiling down to an expectation that the obligation to make such a determination will be performed to ensure that weapons, means or methods of warfare *will neither be developed nor adopted without at least a careful examination of their legality*.

That contention, in turn, begs the question: what *are* the applicable primary norms? While there is widespread agreement on several primary norms, the possible development and employment of AI-related techniques or tools in respect of weapons, means or methods of warfare might nevertheless encounter several disagreements concerning aspects of the sources and/or content of some primary norms. Some of those disagreements stretch back decades (if not longer). Other are relatively new. Such differential approaches as to what constitutes lawful and unlawful conduct prevent normative uniformity and legal universality and thereby preclude the establishment of a comprehensive set of agreed primary legal norms against which all weapons, means or methods of warfare must be reviewed. Consider three examples.

Indiscriminate attacks

First, while there is, to my mind, no reasonable disagreement among States that, in general, indiscriminate attacks are prohibited under IHL/LOAC, some key aspects of that basic principle are currently contested. Take direct participation in hostilities as an example. In general, under IHL/LOAC civilians shall enjoy protection against the effects of hostilities. Certain aspects of those protections—including the so-called immunity from direct attack—might be withdrawn with respect to civilians who take a direct part in hostilities. There seems to be extensive support for the customary principle upon which *Article 51(3)* of AP I is based. (That provision, at least as a matter of treaty law, concerns direct participation of civilians in hostilities in respect of international armed conflicts as defined in that instrument.) Yet, *according* to the *Law of War Manual* (December 2016), the Office the General Counsel of the United States Department of Defense has noted that, at least in its view, that treaty provision, as drafted, does not reflect customary international law in all of its precise aspects.

Applicable legal frameworks

Second, with respect to *applicable legal frameworks*, there is, to my mind, no reasonable disagreement among States that relevant provisions of at least IHL/LOAC must be taken into account in legal reviews of weapons. Meanwhile, some States are considering whether international human rights law (IHRL) provisions must also be taken into account—and, if so, how and to what extent. The United Kingdom, for example, is apparently actively considering this issue. Such an assessment concerning the applicable framework(s) matters in no small part because the content of relevant IHL/LOAC provisions are at least traditionally perceived as tolerating more—indeed, in certain circumstances *much more*, though never unlimited—death, destruction and other harm in comparison to IHRL provisions.

A primary norm concerning AI-related techniques or tools?

Third, there currently seems to be a pivotal disagreement among certain States whether a new o emerging primary norm concerning AI-related techniques or tools and other relevant technologies can, should and/or must be developed. (According to certain scholars and advocates, such a norm might already be discerned.)

Here is where much of the normative debate currently seems to lie in respect of ‘emerging technologies in the area of lethal autonomous weapons systems’ (to use the term from the *title* o the relevant Group of Governmental Experts). On one hand, for some States, such a primary norm might be formulated in conceptual terms drawn, for example, from the August 2018 proposal by Austria, Brazil and Chile to establish a mandate for a new binding international instrument. That proposal *speaks* of 'ensur[ing] meaningful human control over critical function in lethal autonomous weapon systems'. On the other hand, certain other States *argue* that existing IHL/LOAC is sufficient. According to that viewpoint, the 'modernization' or 'adaptation of IHL/LOAC in respect of emerging technologies in the area of lethal autonomous weapons systems is not needed.

16 elements or properties of interest or concern

While recognizing the significance of the disagreement on the existence and/or sources—or, at least, on some precise aspects—of certain primary norms identified above, it remains imperativ for States to adopt robust legal review regimes. With that in mind, it may be of value to enumerate elements or properties of interest or concern that might be salient for the people responsible for conducting legal reviews of weapons, means or methods of warfare involving AI-related techniques or tools to consider.

A few caveats first. The listing order here is not meant to imply a hierarchy. Some of the element or properties might overlap substantively and/or procedurally. Others might stand on their own. Inclusion on the list is not meant to represent a contention that international law does or does not already oblige a State to consider that particular element or property as part of a legal review Nor is the list meant to exhaustively enumerate all possibly relevant considerations—far from it With those caveats in view, here are 16 non-exhaustive assessments concerning elements or properties of interest or concern that might be considered as part of a legal review:

1. **Legal agency:** an assessment concerning the preservation of legal agency of humans—as grounded in international law—in respect of an employment of weapons, means or methods o warfare involving AI-related techniques or tools;
2. **Attributability:** an assessment concerning the preservation of the attributability—at least to a State and to an individual, including, as relevant, a commander—of an employment of weapons, means or methods of warfare involving AI-related techniques or tools;
3. **Explainability:** an assessment concerning the preservation of the explainability of an employment of weapons, means or methods of warfare involving AI-related techniques or tools;
4. **Reconstructability:** an assessment concerning the preservation of the reconstructability—in a nutshell, the capacity to sufficiently piece together the inputs, functions, dependencies and outputs of the computational components adopted, and by whom, in relation to each relevant circumstance of use, encompassing all potential legal consequences thereof—of an employment of weapons, means or methods of warfare involving AI-related techniques or tools both during and after employment (a possible guidepost here might be that such an

employment is capable of being subject to juridical scrutiny, including by a judicial organ);

5. **Proxies:** an assessment whether the computational components—adopted in respect of an employment of weapons, means or methods of warfare involving AI-related techniques or tools—may or may not be permitted to function, in whole or in part, as proxies for any legally relevant characteristics;
6. **Human intent and human knowledge:** an assessment concerning the preservation of human intent and human knowledge—as they pertain to compliance with international law applicabl in relation to armed conflict as regards State responsibility and/or individual (including *criminal*) responsibility—in respect of an employment of weapons, means or methods of warfare involving AI-related techniques or tools;
7. **Normative inversion:** an assessment concerning the preclusion of normative inversion—that is, preventing the computational components from operating in a manner that, for example, assumes that every person may prima facie be directly attacked, thereby functionally rejecting and hence inverting, the general presumption of (protected) civilian status—in respect of an employment of weapons, means or methods of warfare involving AI-related techniques or tools;
8. **Value decisions and normative judgments:** an assessment concerning the reservation of IHL/LOAC-related value decisions and normative judgments only to humans in respect of an employment of weapons, means or methods of warfare involving AI-related techniques or tools;
9. **Ongoing monitoring:** an assessment concerning the feasibility or not of the ongoing monitoring of the operation of the computational components adopted in an employment of weapons, means and methods of warfare involving AI-related techniques or tools;
10. **Deactivation and/or additional review:** an assessment concerning the feasibility or not of the establishment of deactivation thresholds and/or additional review thresholds in respect of an employment of weapons, means or methods of warfare involving AI-related techniques or tools;
11. **Critical safety features:** an assessment concerning the prevention of the continued employment of weapons, means or methods of warfare involving AI-related techniques or tools where a critical safety feature has been degraded;
12. **Improvisation:** an assessment concerning the establishment of sufficient limitations and—as warranted—prohibitions on possible forms of 'improvisation' in relation to an employment o weapons, means or methods of warfare involving AI-related techniques or tools;
13. **Representations:** an assessment concerning the representations reflected in the computational components—in short, the configurations of the models and their features—adopted in respect of an employment of weapons, means or methods of warfare involving AI-related techniques or tools;
14. **Biases:** an assessment concerning the biases capable of arising in relation to the computation

components adopted in respect of an employment of weapons, means or methods of warfare involving AI-related techniques or tools;

15. **Dependencies:** an assessment concerning the dependencies within and between the computational components—and the relationships between those dependencies—adopted in respect of an employment of weapons, means or methods of warfare involving AI-related techniques or tools; and
16. **Predictive maintenance:** an assessment concerning the feasibility or not of the establishment of predictive maintenance—that is, measures aimed at anticipating, forewarning and preventing failures, degradation, or damage with a view to avoiding the need for corrective maintenance—in respect of an employment of weapons, means or methods of warfare involving AI-related techniques or tools.

***

This post is part of the AI blog series, stemming from the December 2018 workshop on Artificial Intelligence at the Frontiers of International Law concerning Armed Conflict held at Harvard Lav School, co-sponsored by the *Harvard Law School Program on International Law and Armed Conflict* the *International Committee of the Red Cross Regional Delegation for the United States and Canada* and the *Stockton Center* for International Law, U.S. Naval War College.

Other blog posts in the series include

- Intro to series and *Expert views on the frontiers of artificial intelligence and conflict*
- Ashley Deeks, *Detaining by algorithm*
- Lorna McGregor, *The need for clear governance frameworks on predictive algorithms in military settings*
- Tess Bridgeman, *The viability of data-reliant predictive systems in armed conflict detention*
- Suresh Venkatasubramanian, *Algorithms and the law: Risk assessment, targeting and cognitive disconnects*
- Li Qiang and Xie Dan, *Legal regulation of AI weapons under international humanitarian law: A Chinese perspective*
- Netta Goussac, *Safety net or tangled web: Legal reviews of AI in weapons and war-fighting*

DISCLAIMER: Posts and discussion on the Humanitarian Law & Policy blog may not be interpreted as positioning the ICRC in any way, nor does the blog's content amount to formal policy or doctrine, unles specifically indicated.