Arms Control and International Security Papers
Volume I | Number 2
April 20, 2020

AI, Human-Machine Interaction, and Autonomous Weapons: Thinking Carefully About Taking "Killer Robots" Seriously

by Christopher A. Ford

The Arms Control and International Security Papers are produced by the Office of the Under Secretary of State for Arms Control and International Security in order to make U.S. State Department policy analysis available in an electronically-accessible format compatible with "social distancing" during the COVID-19 crisis.

AI, Human-Machine Interaction, and Autonomous Weapons: Thinking Carefully About Taking "Killer Robots" Seriously

by Christopher A. Ford[1]

This second in the T series of papers offers thoughts on the public policy challenges presented by the prospect of Lethal Autonomous Weapons Systems (LAWS). In this paper, Assistant Secretary Ford offers his perspective upon these issues, urging readers not to be seduced by sensationalized simplifications, and calling for careful, sustained attention to the complexities they raised – such as through the work already being done by the LAWS Group of Governmental Experts.

It is hard not to be impressed by all the work being done and innovative approaches being pursued in the realms of Artificial Intelligence (AI) and human-machine interaction. Such initiatives have the promise of enriching human life in a vast number of ways, and it is an extraordinarily exciting field to follow.

At the U.S. State Department's Bureau of International Security and Nonproliferation, however, we are concerned with protecting against the potential "dark side" of such innovation – that is, with ensuring that such creations are not manipulated into doing the dirty work of despots or violent non-state actors to harm the innocent, silence the weak, or destabilize the global balance of power. This too, I have come to learn over my years in public service, is no small or easy task. In order for the public policy community to get these answers right, and that we must, we must ask tough questions and not settle on purported solutions just because they seem simple or easy. Like the scientific and technical fields with which they interact, such policy challenges are multi-disciplinary and complex. In order to create a lasting and effective policy for dealing with such dynamic technologies, international policy makers and technology experts will need to work together as perhaps never before.

I. Temptations of Sensationalism

For something as new and thought-provoking as the technological, ethical, legal, and political implications of Lethal Autonomous Weapons Systems (LAWS), the subject has wasted no time in acquiring its shopworn clichés and armchair experts. Who, at this point, doesn't approach a discussion of LAWS without thinking, even inadvertently, of the villainous "Skynet" artificial intelligence system of movie fiction and the robot-assassins it dispatches against the noble but hard-pressed remnants of humanity in Hollywood's Terminator franchise?

There is, of course, a reason that public policy debates over LAWS have to struggle through endless Terminator tropes: activists concerned about the

[1] *Dr. Ford serves as U.S. Assistant Secretary of State for International Security and Nonproliferation, and is additionally performing the Duties of the Under Secretary for Arms Control and International Security. He previously served as Special Assistant to the President and Senior Director for Weapons of Mass Destruction and Counterproliferation on the U.S. National Security Council staff.*

possibility of LAWS have built their public messaging around evocative "Skynet" imagery of "killer robots" precisely because this presses the kind of emotive buttons that tend to presuppose conclusions and short-circuit debate. If that's what is meant by "killer robots," who wouldn't be opposed to them?

The issue certainly seems simple at first glance, anyway. The self-described "Campaign to Stop Killer Robots," for instance, decries a future in which "fully autonomous weapons" would "decide who lives and dies, without further intervention." This, the campaign says, "crosses a moral threshold" because machines "lack the inherently human characteristics such as compassion that are necessary to make complex ethical choices" and "lack the human judgment necessary to evaluate the proportionality of an attack, distinguish civilian from combatant, and abide by other core principles of the laws of war." As a result, it is declared, such machines "would make tragic mistakes with unanticipated consequences" – errors for which it is "unclear" who could be held responsible.

These are strong claims, but unfortunately, we know full well how alarmist and demonizing rhetoric can – as it is so often designed to do – skew policy debate in reflexive and counterproductive directions. Such language is a very powerful tool. To be sure, as Cardinal John Henry Newman once observed, men will die for a dogma who will scarcely stir for a mere conclusion. When it comes to complicated and fraught things like how to deal with technological change in international security policy, however, I hope you'll agree with me that it is best to avoid dogma – and to try to hold out for actual conclusions.

In fairness, of course, it is also quite clear that there are some very real issues in play – and that the emergence of a functioning LAWS, at such point as it materializes, could present all of us with important ethical, moral, and governance challenges. Worrying about LAWS is not something that should be left entirely to dystopic science fiction. There are real public policy questions here.

II. Emerging Challenges

To take just one example, the People's Republic of China (PRC) already invests huge sums in using artificial intelligence (AI) tools as instruments of domestic oppression, and it explicitly envisions military AI to be the key to revolutionizing Beijing's military potential vis-à-vis the United States. Indeed, going far beyond simply talking about autonomous individual weapons systems, a senior executive at one of China's largest defense companies has claimed that AI will be the "brain" of future warfare – with an "AI cluster" taking over from actual humans, for example, in the national command structure.

Nor is this senior executive alone among Chinese thinkers in predicting that AI will prove to be a crucial element of the next Revolution in Military Affairs. According to Ministry of National Defense officials, for instance, the PRC must "strengthen the research of artificial intelligence technology in the military field," in order to allow the PRC to "capture the 'new commanding heights' of the future battlefield, and ensure that our army is built into a world-class military at an early date." Experts at the Chinese Academy of Engineering have predicted that artificial intelligence will be the most important military-civilian dual-use technology in the coming decades.

Even allowing for other strategic powers' ambitious goals, there is unquestionably something worth taking seriously here. This is of particular interest to the policy community and can safely argue it should be of concern to those with the skillset and knowledge to create such systems. My point is not to single out the PRC per se, but rather to make it clear that we are probably only at the beginning of mankind's exploration of the intersection between AI and warfighting – of which the development of LAWS is merely one potential issue.

The United States has recognized from the beginning that the foundation for AI adoption must be guided by ethical foundations deeply rooted in our nation's values and respect for rule of law. Although technology changes, the U.S. commitment to the Law of War and ethical behavior does not. That is why we are leveraging U.S. AI innovations to build solutions that are aligned with our laws and values. For the United States, the Law of War – and basic questions of safety – are key considerations with any new capability, from first step of requirements development through the last day of deployment.

In sharp contrast to the opacity so far displayed by PRC and Russian officials on such subjects, the U.S. Department of Defense (DoD) is promoting thoughtful, responsible, and human-centric adoption of AI by investing in AI systems that are resilient, reliable, and secure. For this reason, the Defense Department asked the Defense Innovation Board (DIB) to propose AI ethical principles for the Department of Defense.

The DIB conducted – in a fully transparent manner – a 15-month study that included consultation with many leading AI and technical experts, current and former DoD leaders, and the American public. On February 25, 2020, Secretary of Defense Esper adopted the five ethical principles proposed by the DIB for the DoD.

The five principles – under which AI in defense must be responsible, equitable, traceable, reliable, and governable – apply to both combat and non-combat AI technologies used by the DoD. The U.S. system of democratic values and transparency, which led to the development of the DoD's AI ethics principles, provides a framework for likeminded nations to follow as they look to develop their own AI principles.

III. LAWS and Their Future

But whatever the future of AI in general, LAWS debates only concern the modest subset of AI-related issues that pertain to potential autonomous features and functions in weapons. LAWS discussions are frequently conflated with broader AI debates, but they are not the same, and such conflation makes it needlessly difficult to think about either one.

Identifying what elements to focus on is not an easy task in this extraordinarily complicated and fast-moving arena, full of potential distractions. Most of us catch the "Skynet" references, and some of us who are old enough will remember the rogue computer HAL from Stanley's Kubrick's masterful film 2001: A Space Odyssey, but it is also true that pop culture predictions of the future have a notably poor track record. Having discovered that we don't actually now live in George Jetson's world – or the world of chauffeured craft swimming through the air to and from the Paris Opera depicted in that marvelous illustration by the 19th-Century French futurist Albert Robida – we should have more intellectual humility than to think we can understand all that much about our technological future.

If we can muster that humility, however, I do believe that there is real value exploring the public policy challenges potentially presented by LAWS. We just need to approach it with care, rather than just reflex, as indeed such questions deserve.

As we consider these matters, therefore, I would like to offer some thoughts on the issues that seem to be the most consequential as we continue to work through the challenges posed by autonomous features and functions in weapons.

Problems with the term "meaningful human control"

Some in the NGO community have said this is really a question of "meaningful human control." This is one of those phrases that seems very simple until one thinks about it carefully. In thinking about this challenge from the perspective of future weaponry, it's useful to remember the degree to which this may not be a completely binary question – "human" versus "no human" – but instead something at least a little more like a continuum.

It's already long been the case, for instance, that some weapons systems incorporate autonomous features and functions – though they do not make their "own" independent decisions but rather respond to pre-established criteria pursuant to rules programmed into them by humans. A deep-sea mine, for instance, might identify a particular hydrophonic signature associated with an enemy submarine, or an anti-tank mine might detect the clank of steel treads. An artillery round or air-delivered submunition might scan the ground as it falls, employing its millimeter-wave sensor to pick up the shape of a tank's armored turret before detonating its self-forging fragment warhead, or a drone might pick up the distinctive electronic emanations of an enemy target-acquisition radar before flying itself down that beam to destroy that emitter. Or a "Close In Weapons System" Gatling gun – the U.S. Navy's so-called CIWS, or "sea-whizz" system – might be set to "automatic" so that it is able to shoot incoming cruise missiles out of the air faster than a human sailor could respond.

Such things have long been realities, but the sky hardly fell upon their adoption, and we are hardly now on an inexorable slippery slope to "Skynet." In fact, I expect that many can see positive utility in these advanced technologies for limiting unintended harm to civilians.

Furthermore, it is already – and, I think, increasingly – the case that for major weapons systems, human-machine interaction in the use of the weapon could actually be positive in increasing the degree to which human intent is precisely effectuated in the use of force. To be sure, the pilot in a first-rate combat aircraft still makes the ultimate "fire-versus-refrain" decisions, but the data upon which such decisions are made is already highly dependent on automated software interfaces that characterize, sort, interpret, and prioritize the output of a huge range of sensors more precisely and more efficiently than any human could do, before the pilot knows anything. Without the output of those sensors –

and the "judgment" employed by the software that determines what to tell the pilot so that combat decisions can be made – the individual pilot would have to make his or her decision about ultimate use of force with very little relevant information at all.

An individual human being could not possibly integrate all the incoming information in real time – or even sense it in the first place, what with radar, infra-red, radiofrequency signals, and data-fusion from a broad network of other sources all flowing into the cockpit, all being machine-processed and fused in real-time, and often concerning candidate targets far enough away that they cannot be personally seen by the pilot in any way. Increasingly, the modern combat pilot necessarily understands his or her operational environment through software of extraordinary complexity. Yes, the human makes the final call, but the pilot is hugely challenged or even helpless without machine intermediation, unable personally to vouch for what is "really" there without trusting the output of machines that he or she did not program in evaluating information that he or she may not directly perceive, in order to make the best and most considered decision possible.

Given that I'd wager that there are certain pilot-interface outputs that would quite invariably result in a macro decision to fire a weapon – something perhaps more or less equating to a signal that "that enemy fighter just locked on to you, is about to fire, and will kill you unless you fire first" – how should we consider the human-machine interaction at the time of "trigger pull"? It is certainly true that for my hypothetical pilot, it is other humans who have programmed the computer to look for certain complex signatures that are identified as threats. Yet, we are comfortable with the pilot firing a weapon in response to information generated by a machine at least in part because of the significant human judgment and expertise that was involved in programming the software and designing the sensors that enable the jet's computer systems to identify military objectives in the pilot's operating environment.

This suggests an important lesson. In many circumstances, it may be the degree of human judgment that is exercised during the development and deployment of a weapon, rather than the degree of human control over the weapon at any given moment that will be critical to ensuring compliance with International Humanitarian Law (IHL).

The importance of context

Perhaps spurred by periodic media stories about drone strikes against terrorist targets in the Middle East – albeit ones undertaken by human pilots via remote control architectures – those who fret about Terminator-style robot assassins seem usually to have in mind contexts in which a machine aims to pick out a combatant from a civilian standing nearby. The implication seems to be that this would inevitably be happening in and among all the ordinary moving pieces of life in the civilian world. (The Campaign to Stop Killer Robots, for instance, warns that "no one would be safe"!)

But context does matter, and it's worth considering whether there are situations in which autonomous features and functions might not be quite so frightening. To begin with, of course, some autonomous features and functions – even deep autonomy, rather than merely automation – might actually be quite beneficial in some areas, with machine learning being used to enhance our everyday lives and perhaps also providing important advantages even in non-lethal military applications such as intelligence and logistics.

And even when it comes to lethal actions, the context and environment in which a weapon system is to be employed must surely matter a great deal. I mentioned the CIWS above, and it's a good example of a system with an autonomous feature that is turned on to "automatic" mode only by a human, and only in a very specific context of high-intensity threat at sea – presumably when there aren't aerial targets around other than incoming enemy aircraft or cruise missiles. CIWS has been an operational reality for decades, and people don't seem particularly unnerved by automaticity in that degree and in that context.

Could one imagine other situations in which autonomous features or functions – with systems being fixed to unambiguously military signatures and only unleashed when a human decides that certain factual predicates have been satisfied – would not be terribly problematic? Perhaps in hunting for armored vehicle silhouettes above a massive tank battle taking place in the desert? Or in high-intensity air battles over the front line in a conventional war between "near-peer" adversaries where there would be essentially no civilian activity at all? Or in defending against incoming ballistic missiles, or where underwater vehicles duel each other at sea? In the right circumstances, it strikes me that lethal autonomy might well be quite defensible and appropriate.

How precisely does one define the problem?

It may also be useful to think a little more carefully about the fundamental terms of the debate. To what degree, for instance, is the crux of the problem the claim that humans inherently make more accurate decisions than machines? While the unresolved problem of data and algorithmic bias are of real concern when evaluating weapons with autonomous features and functions, we cannot escape the reality that humanity's cognitive biases might make us ill suited, by the same measure, to be more accurate at some decisions than machines.

As I noted, the Campaign to Stop Killer Robots argues that machines "lack the human judgment necessary to evaluate the proportionality of an attack, distinguish civilian from combatant, and abide by other core principles of the laws of war." Such a claim, however, should surely have to be defended, rather than just asserted. Even if LAWS never replaces human judgement, we know that our own judgment and intuitions can be quite flawed, and we should not assume that our collective understanding of how human judgement can and should interact with machines will not evolve. When and how such judgment is exercised in the future may look very different than it does today – such as in the proportion of decisions we choose to augment by the use of algorithmic tools precisely in the interests of accuracy and reliability.

But this would hardly equate to lawlessness, or any lack of accountability. IHL obligations are applicable regardless of the type of weapon being employed, as is accountability for adhering to those obligations on humans, whether a LAWS is in the picture or not. Humans and States at all times remain responsible in accordance with applicable law for their decisions, whether in the development and use of a LAWS or not. A debate on accountability is quite different than one focused on functionality.

But more prosaically and less theologically, if the claim is simply that machines cannot make correct operational decisions, that's a different situation entirely. That is an empirical question to which one could imagine actual testing could provide an answer.

Is the solution implementable?

Humanitarian implications aside, there are those advocating for a ban on LAWS for international security concerns. Some states warn of a coming arms race, the likes of which have not been seen since the Cold War. Others, primarily those of the Non-Aligned Movement, warn of a world wherein the haves will use autonomous or AI-augmented tools force their wills upon the have-nots, leading to an unstable shift in the global balance of power.

Well, maybe. We have all heard the saying "hope is not a strategy," and it's true enough. But I would also submit that raw fear – however powerful a motivation it may be – is not a sound basis for strategy either, particularly because it's very good at short-circuiting careful analysis. Moreover, even if there's unquestionably a pressing set of questions here, we should ask ourselves how sure we are that there really is actually a solution available to us – at least one beyond emphasizing, as we and our likeminded Western friends and allies already do, the need for careful law-of-war reviews prior to the deployment of any new weapons system and respect for international law in their use?

At least in the United States, such weapon systems already undergo careful pre-deployment weapons reviews in order to ensure that they can be used consistent with obligations under international humanitarian law. U.S. reviews consider a broad range of factors, including how the weapon will be deployed, its intended operating environment, the concept of operations underlying the system, the tactics, techniques, and procedures (TTPs) developed for its employment, and the rules of engagement (ROEs) envisioned for it. Should any such potential weapon system fail review, it would not be deployed, or it would be redesigned until it passed muster.

I am aware of no persuasive evidence as to why weapons reviews by definition are unable to address the challenges of LAWS or other future weapons systems. We certainly need to be encouraging states to improve their practices with regard to these reviews, especially countries such as the PRC and Russia, which seem to be moving as fast as they can into the development of AI-related military applications. We also need to ensure that all weapons can be used consistent with international humanitarian law obligations – not just LAWS.

Beyond such common-sense approaches, however, the complexities of this environment resist simplistic solutions. Given the fast-moving nature of the technology in question and the protean character of this new field – and in light of some of the complexities to which I've tried to draw your attention – it is hard to have confidence that the category of a "lethal autonomous weapon system" is definable enough to be "ban-able" in the first place. And even if you could ban it, how effective or enforceable would such a ban actually be?

How, for instance, would you verify compliance with whatever specific prohibition you ended up deciding was needed? I suppose you could imagine a weapon system the control architecture of which was somehow hardwired to be hermetically sealed-off from the outside world and hence physically incapable of non-autonomy, and hence hypothetically prohibited. Where systems have two-way communications links to elsewhere, however, the key to evaluating whether or not there is inescapably a "human in the loop" would presumably lie in fully understanding the system's software. And that certainly seems to present a difficult verification challenge.

I have personally heard it suggested that one might establish a global software declaration process for all weapons systems, in which source code would be provided to some kind of international verification authority analogous to the International Atomic Energy Agency (IAEA) or the Organization for the Prohibition of Chemical Weapons (OPCW), where expert computer scientists would vet each software package and certify whether or not the system is capable of autonomous lethality. But could one ever imagine that countries would agree to turn over all the source code for their cutting-edge weapons systems?

And even if countries did make such disclosures, could such an authority possibly vet them properly or keep said technologies safe from theft, manipulation, sabotage, or proliferation? And how on Earth would you know whether or not any given declarer had secretly held something back? Or that it hadn't just modified the software right after getting an international seal of approval? Or simply used another computer algorithm to provide the final decision to be beamed back into the device in response to a query about whether or not to shoot? The "international verification" notion seems to me an entirely unworkable idea, incapable of functioning effectively, or being at all trustworthy.

And if one threw up one's hands in frustration at such an unworkable idea and opted simply for a convention that merely banned LAWS but lacked any kind of transparency or verification protocols, who would be constrained by that scheme? I would venture to say that some scrupulous governments who chose to become party to such a convention certainly would be so constrained – just as we in the United States are constrained by the terms of the Biological and Toxin Weapons Convention (BTWC), even though it has never had a verification protocol.

But it also seems safe to say that some governments wouldn't feel constrained by a LAWS ban without a verification mechanism, either out of lack of respect for international law, or by simply declining to join the convention. After all, both Russia and Syria clearly maintain, and have even used, prohibited chemical weapons even though they are States Party to the Chemical Weapons Convention (CWC) – and that treaty actually does have declaration and verification rules, as well as a specialized international organization to help implement them.

A purely normative LAWS ban, therefore, seems likely to have an asymmetric result. The most unscrupulous builders and users of high-technology weaponry would be the ones least likely to respect obligations undertaken, and many of the most sophisticated state users, including responsible ones, would simply decline to participate in the convention for all the reasons I have already outlined.

Nor, I suspect, could we rely upon NGO "killer robots" activists to mobilize civil society pressures to bring things under control, for it is precisely the governments most immune to such pressures – such as the high-technology police state of the Chinese Communist Party, or the autocracy of Vladimir Putin in Russia – who would seem most likely to build lethal autonomous systems in violation of a ban. In other words, the rules would work best where you needed them the least and work least where you needed them the most (i.e., in constraining scofflaw states such as Russia and China keen to obtain military advantages versus the West).

As I have also observed of the so-called nuclear weapons "ban" treaty, the asymmetric impact of prohibition activism upon free, democratic states thus suggests the possibility that to the extent that such efforts actually succeed, they may risk creating a dynamic of de facto disarmament by the world's democracies vis-a-vis authoritarian states. As you might imagine, I'm far from convinced that that's a good idea – especially given the degree that Chinese strategic thinkers openly write about how they see their road to mid-21st-Century military dominance being driven in large part by pioneering work in AI-driven, "intelligentized" weapons systems and command-and-control architectures.

IV. The LAWS GGE Process

In fairness, it is not quite right to suggest a LAWS ban cannot be a policy option, as I suppose it certainly could be an option for some states. I posit, however, that it would not be an effective option, and that it could itself present some very formidable problems – especially as applied to the powers one would most like to see forced to behave ethically. Instead of fixating upon such simplistic answers, therefore, I urge the international community to roll up its sleeves, ask the tough questions, and continue to improve our understanding of the issues before jumping into the deep end of purported policy "solutions." Through deeper and more constructive discussions with experts, policy makers may better gauge the myriad of connected issues instead of creating knee-jerk policy prescriptions that feed reactionary fears.

On each of these key matters, there may be legitimate differences of view. But one thing is clear: these are issues that cannot be resolved in politically-charged bumper sticker-level debates over "killer robots."

Fortunately there is already a process in the international community aimed at addressing the complexities of these issues, with engagement from both states and NGOs and other external actors, which is making progress as we speak. The Group of Governmental Experts (GGE) on LAWS, convening under the auspices of the UN Convention on Certain Conventional Weapons, is working hard to make substantive progress on these complex questions.

Critics of this process, I think, do not appreciate how uniquely well-suited to this debate the GGE has been and continues to be. It is a standing forum that meets both formally and informally for several weeks each year, with a mandate devoted exclusively to LAWS-related issues as they pertain to IHL. It's worth emphasizing how unusual and promising this is. Unfortunately, proposals for even standing up such a forum in other IHL contexts – for instance, to discuss strengthening protections for IHL – have foundered. In many cases, states cannot even agree to meet regularly to discuss complicated issues like this.

But the LAWS GGE does all this as a matter of course, as part of the CCW framework. This in and of itself has value. Getting all of these actors – from the United States to Costa Rica, from Russia and the PRC to Austria and Brazil, and everyone in between – into a room together to hash through these issues and speak frankly about them ensures that the issue continues to get the attention from governments that it deserves, all under the watchful eye of responsible civil society.

Calls to have the discussion moved elsewhere, or undue impatience with these deliberations, could make the problem worse, not better. At present a de-politicized diplomatic process is grappling thoughtfully with the core issues that states will have to face. Activists should be thankful these matters are being taken seriously, and not push for moves that would undermine the good work that's going on.

Jumping into sweeping policy prescriptions without a proper diagnosis of the problem or understanding of its complexities would be folly indeed. Diplomacy takes time and developing a usefully comprehensive understanding of such difficult legal and technical issues takes time. The GGE should be given enough time to do the heavy lifting of discourse, debate, sharing of best practices, and understanding.

The GGE, in fact, has already made real progress – including recommending the adoption of 11 Guiding Principles for the Responsible Use and Development of LAWS to the CCW meeting of High Contracting Parties last year. These are real, substantive, consensus principles adopted through genuine debate and discussion, grounded in respect for international humanitarian law. One cannot say with a straight face that this is not progress.

V. Conclusion

If we in government are to be responsible stewards of the security and safety of the civilian populations entrusted to our care, we owe these issues care and consideration. I am not arguing that these questions are necessarily unanswerable, nor that the answer is necessarily that LAWS should not or cannot somehow be addressed by the international community. I merely contend that answering these questions is difficult, that answering them in some fashion is necessary, and that we already have a place well suited to seeking answers to these questions in the international community at the LAWS GGE.

I can assure you that no one wants to see the emergence of something like "Skynet" any less than I do. I do hope, however, that this paper has helped illuminate the complex nature of the topic. LAWS is an important that deserves being taken seriously by policy makers and technologists alike, rather than addressed as a policy reflex.

* * *

Arms Control and International Security Papers

The Arms Control and International Security Papers are produced by the Office of the Under Secretary of State for Arms Control and International Security in order to make U.S. State Department policy analysis available in an electronically-accessible format compatible with “social distancing” during the COVID-19 crisis.