It Takes a Village: The Shared Responsibility of 'Raising' an Autonomous Weapon

Amritha Jayanti and Shahar Avin

Submitted to and Presented at RSIS Military Transformations Programme Workshop
Co-organised with the Norwegian Institute of Defence Studies (IFS) &
the Institute for Security Policy at Kiel University

"Artificial Intelligence, Robotics, and the Future of Defence:
The Role of Ethics, Norms, and Governance"

Abstract

Expectations around future capabilities of lethal autonomous weapons systems (LAWS) have raised concerns for military risks, ethics, and accountability. The U.K.'s position, as presented among various international voices at the UN's Convention on Certain Conventional Weapons (CCW) meetings, has attempted to address these concerns through a focused look at the weapons review process, human-machine teaming or "meaningful human control" (see e.g. JCN1/18), and the ability of autonomous systems to adhere to the Rules of Engagement. Further, the U.K. has stated that the existing governance structures—both domestic and international—around weapons systems are sufficient in dealing with any concerns around the development, deployment, and accountability for emerging LAWS; there is no need for novel agreements on the control of these weapons systems. In an effort to better understand and test the U.K. position on LAWS, the Centre for the Study of Existential Risk has run a research project in which we interviewed experts in multiple relevant organisations, structured around a mock parliamentary inquiry of a hypothetical LAWS-related civilian death. The responses to this scenario have highlighted different, sometimes complementary and sometimes contradicting, conceptions of future systems, challenges, and accountability measures. They have provided rich "on the ground" perspectives, while also highlighting key gaps that should be addressed by every military that is considering acquisition and deployment of autonomous and semi-autonomous weapon systems.

Introduction

With the increasing integration of digital capabilities in military technologies, many spheres of the public--from academics to policymakers to legal experts to nonprofit organizations--have voiced concerns about the governance of more "autonomous" weapons systems. The question of whether autonomous weapons systems pose novel risks to the integrity of governance, especially as it depends so heavily on the concept of human control, responsibility, and accountability, has become central to the conversations.

The United Kingdom (U.K.) has posited that lethal autonomous weapons (LAWS), in their current and foreseeable form, do not introduce weaknesses in governance; existing governance and accountability systems are sufficient to manage the research, development, and deployment of such systems and the most important thing we can do is focus on improving our human-machine teaming. Our research project seeks to test this theory by asking: With the introduction of increasingly autonomous agents in war (lethal autonomous weapons/LAWS), are the current governance structures (legal, organizational, social) in fact sufficient for retaining appropriate governance and accountability in the U.K. MoD? By attempting to confront strengths and weaknesses of existing governance systems as they apply to LAWS through a mock parliamentary inquiry, the project uncovers opportunities for governance improvements within Western military systems, such as the U.K.

Background

Computers and algorithms are playing a larger and larger role in modern warfare. Starting around 2007 with writings by Noel Sharkey, a roboticist who heavily discusses the reality of robot war, members of the research community have argued that the transition in military technology research, development, and acquisition to more autonomous systems has significant, yet largely ignored, moral implications for how effectively states can implement the laws of war.[1] Segments of this community are concerned with the ethics of decision making by autonomous systems, while other segments believe the key concern is regarding accountability: how responsibility for mistakes is to be allocated and punished. Other concerns raised in this context, e.g. the effects of autonomous weapon systems on the likelihood of war, proliferation to non-state actors, and strategic stability, are beyond the scope of this brief, though they also merit attention.

U.K. Position on LAWS

The United Kingdom's representatives at the UN Group of Governmental Experts (GGE) on Lethal Autonomous Weapon Systems (LAWS) have stated that the U.K. believes the discussions should "continue to focus on the need for human control over weapon systems and that the GGE should seek agreement on what elements of control over weapon systems should be retained by humans."[2] The U.K., along with other actors, such as the United States, believe that a full ban on LAWS could be counterproductive, and that there are existing governance structures in place to provide appropriate oversight over the research, development, and deployment of automated weapons systems:

> …[T]he U.K. already operates a robust framework for ensuring that any new weapon or weapon system can be used legally under IHL. New weapons and weapons systems are conceived and created to fulfil a specific requirement and are tested for compliance with international law obligations at several stages of development.[3]

The U.K. is also interested in a "technology-agnostic" focus on human control because it believes that it will "enable particular attention to be paid to the key elements influencing legal, ethical and technical considerations of LAWS, as opposed to "debated definitions and characteristics" which, ultimately, may "never reach consensus." The position emphasizes that taking a "human-centric, through-life" approach would enable human control to be considered at various stages and from multiple perspectives. This includes across all Defense Lines of Development, the acquisition of weapons systems, and their deployment and operation. It is the U.K.'s position that the existing institutional infrastructure builds-in accountability measures throughout the weapon system lifecycle.

Methodology

In order to stress test the governance and accountability structures that exist for U.K. weapon systems, and how they would apply to LAWS, we developed a hypothetical future scenario in which a U.K. LAWS kills a civilian during an extraction mission in Egypt. In order to ensure a level of feasibility and accuracy of construction, the scenario was built based on a wargaming scenario publicly published by RAND.[4] We then ran a facilitated role-play exercise based on our modified scenario with an initial group of Cambridge-based experts. With their feedback and the lessons from the role-play, we developed the final version of the scenario which we then used in the research study (see Appendix).

[1] Carpenter, C. (2014). From "Stop the Robot Wars!" to "Ban Killer Robots." *Lost Causes*, 88–121. doi: 10.7591/cornell/9780801448850.003.0005

[2] Human Machine Touchpoints: The United Kingdom's perspective on human control over weapon development and targeting cycles, Human Machine Touchpoints: The United Kingdom's perspective on human control over weapon development and targeting cycles (2018).

[3] Ibid.

[4] Khalilzad, Z., & Lesser, I. O. (1998). *Selected Scenarios* from *Sources of Conflict in the 21st Century* (pp. 317–318). RAND.

This final iteration of the LAWS scenario was used to run a mock U.K. parliamentary inquiry through which we interviewed 18 experts across various areas of expertise, including (but not limited to) U.K. military strategy, military procurement, weapons development, international humanitarian law, domestic military law, military ethics, and robotics.

The interviews ranged from 45 to 120 minutes and explored a variety of questions regarding the case. The main objective of the interviews was to catalyze a meaningful discussion around what information the experts deemed important and necessary in order to decide who should be held accountable in the aftermath of this scenario. A sample of the questions asked include:

- Who holds the burden of accountability and responsibility?
- What explanations and justifications for actions are needed?
- What information is necessary to come to a conclusion about the burden of accountability?
- Are there any foreseeable gaps because of the autonomy of the weapons systems?

The responses and dialogue of these 18 interviews were then reviewed and synthesized in order to develop a landscape of strengths and weaknesses of the current governance and accountability schemes related to U.K. institutions as they relate to LAWS, as well as recommendations on addressing any identified weaknesses. The full report is under preparation, but we are happy to share our preliminary key findings and recommendations below.

Key Findings

The main takeaway from the "inquiry," from both a legal and organizational standpoint, was that assessing accountability is in the details. This contrasts with what we perceive as a dominant narrative of "meaningful human control," which focuses mainly on human control, and the design of that interaction, at the point of final targeting action. The disconnect between the accountability across a weapon's lifetime and the focus on final targeting decision was observed throughout the various expert interviews. "Meaningful human control" has become the idée fixe of domestic and international conversations for regulation of LAWS but it disadvantageously provides a limited lens through which most experts and relevant personnel think about accountability.

To contrast this heavily focused narrative, the interviews have highlighted a whole range of intervention points, where humans are expected to, and should be supported in making decisions that enable legal, safe, and ethical weapon systems. These are arguably points that should be considered in "meaningful human control." These include, but are not limited to:

- **Establishment of military need:** defining military necessity for research, development, and/or procurement; choice of technological approach based on political and strategic motivations. *(Main related stakeholders: U.K. MoD; U.K. Defense Equipment and Support (DE&S); Private military contracting companies, such as BAE Systems, Qinetiq, General Dynamics.)*
- **Technical capabilities and design:** trade-offs between general applicability and tailored, specific solutions with high efficacy and guarantees on performance; awareness, training, and foreseeability of contextual factors about intended use situations that may affect the performance of the weapon system; documentation and communication of known limitations and failure modes of the system design. *(Main related stakeholders: Private military contracting companies, such as BAE Systems, Qinetiq, General Dynamics; U.K. Defense Science and Technology, U.K. Defense and Security Analysis Division)*
- **Human-computer interaction design:** choices of what data to include and what data to exclude; trade-offs between clarity and comprehensiveness; level of technical information communicated; parallel communication channels: to operator in/on the loop, to command centres further from the field, logs for future technical analysis or legal investigation. *(Main related stakeholders: Private military contracting companies, such as BAE Systems, Qinetiq, General Dynamics; U.K. Defense Science and Technology; U.K. Defense and Security Analysis Division; U.K. MoD military personnel - human operators)*
- **Weapons testing:** choice of parameters to be evaluated, frequency of evaluation, conditions under which to evaluate; simulation of adversaries and unexpected situations in the evaluation phase; evaluation of HCI in extreme conditions; evaluation of the human-machine team. *(Main*

related stakeholders: Private military contracting companies, such as BAE Systems, Qinetiq, General Dynamics; U.K. DE&S; U.K. MoD military personnel - human operators)

- **Procurement:** robust Article 36 review; assessment of operational gaps, and trading-off operational capability with risks; trade-off between cost effectiveness and performance of weapons systems; documentation and communication of trade-offs so they can be re-evaluated as context or technology changes; number and type of systems; provisioning of training and guidance; provisioning for maintenance. *(Main related stakeholders: U.K. DE&S; Article 36 convened expert assessment group; Private military contracting companies, such as BAE Systems, Qinetiq, General Dynamics)*
- **Weapons deployment:** informing commanders about the capabilities and limitations of the system, of their track record in similar situations, of novel parameters of the new situation; establishing and training for appropriate pre-deployment testing schemes to capture any vulnerabilities or "bugs" with any specific weapons system; checking for readiness of troops to operate and maintain systems in the arena; expected response of non-combatants to the presence of the weapon system. *(Main related stakeholders: U.K. MoD commanding officers; U.K. MoD military personnel -- human operators)*
- **Weapons engagement:** awareness of limiting contextual factors, need to maintain operator awareness and contextual knowledge; handover of control between operators during an operation. *(Main related stakeholders: U.K. MoD military personnel -- human operators)*
- **Performance feedback:** ensuring a meaningful feedback process to guarantee process improvement, reporting of faulty actions, communicating sub-par human-machine techniques and capabilities, and more. *(Main related stakeholders: U.K. MoD military personnel -- human operators; U.K. MoD commanding officers; U.K. DE&S; Private military contracting companies, such as BAE Systems, Qinetiq, General Dynamics)*

Recommendations

Dialogue Shift: Emphasizing Control Chain and Shared Responsibility

The prioritization of "meaningful human control" for LAWS-related risk mitigation and governance anchors the scope of control points around final targeting decisions. The narrative implies that this is the main area of control that we want to manage, focus, and improve on in order to ensure that the weapons systems we are deploying are still acting with the intent and direction of human operators. Although this is an important component of ensuring thoughtful and safe autonomous weapons systems, this is only a fraction of the scope of control points. In order for us to acknowledge the other points of control throughout the research, development, procurement, and deployment of LAWS, we need to be inclusive in our dialogue about these other points of human control.

Distribution of Knowledge: Personnel Training

Training everyone who touches the research, development, deployment, etc. of LAWS on international humanitarian law, robot ethics, legality of development, responsibility schemes, and more, would contribute to a more holistic approach to responsibility and accountability, and, at its best, can contribute to a culture that actively seeks to minimise and eliminate responsibility gaps through a collaborative governance system.[5] This distribution of understanding around governance could provide a better landscape for accountability through heightened understanding of how to contextualize technical decisions. Further, it can provide an effective, granular method for protecting against various levels of procedural deterioration. With shifting geopolitical pressures, as well as various financial incentives, there could easily be a deterioration of standards and best practices. A collaborative governance scheme that is based on a distributed understanding of standards, military scope, international norms, and more, can provide components of a meaningful and robust governance plan

[5] Ansell, C. (2012). Collaborative Governance. Retrieved from https://oxfordindex.oup.com/view/10.1093/oxfordhb/9780199560530.013.0035.

for LAWS. This distribution of knowledge, though, must be coupled with techniques for reporting and transparency of procedure to be effective.

Acknowledging the Politics of Technical Decision Making/Design Specifications

"Meaningful human control," through its dialogue anchoring, also puts a heavy burden on the technical components of design decisions, such as best practices for human-computer interactions. The politics of quantification in technical decision systems for autonomous systems should not be undervalued. The way any autonomous system decides what actions to take and what information to show is a highly political decision, especially in the context of war. It is important to understand which parts of the design process are more political than they are technical, who should be involved in those decisions, and how to account for those decisions in the scope of research and development (to inform a proper, comprehensive collective responsibility scheme).

Appendix

Opposed Evacuation from a Collapsing Egypt

Political-Military Context

Egypt is convulsed by internal instability, with the Egyptian government under siege from well-organized and well-financed anti-Western Islamic political groups. The government has not yet fallen, but political control has broken down, and there is a strong likelihood that the government will indeed collapse. There are large numbers of running battles between government forces and the opposition, with the level and frequency of violence steadily escalating.

U.K. citizens are being expressly targeted by the opposition, and many of the 17,000 or so Britons in Egypt—along with other Westerners—have taken refuge in the major urban areas. The Egyptian military has so far proved largely loyal to the government, but some troops—including army, air force, and naval units—have sided with the Islamic opposition, and the allegiances of many other elements are unclear. At least one crack armor brigade has joined the opposition en masse and is operating in the Cairo area. Security at airports and seaports is breaking down, with anti government elements in control of some. Opposition leaders have indicated that they will oppose any attempt to evacuate Western citizens with "all available means and the assured assistance of Allah."

With the small amount of remaining government, an expedited bilateral Status of Forces Agreement (SOFA) between the U.K. and Egypt has been adopted that discredits/overrides local law (citing the quick deterioration of its human rights-compatible justice system).

U.K. Objectives

Approximately 20,000 to 23,000 U.K., other Western, and friendly Egyptian personnel are now in direct danger as the host government nears collapse. These people are in need of rapid (48–96 hours) evacuation and rescue.

U.K. military objectives are to:

- Secure necessary aerial and seaports of embarkation to support evacuation operations
- Establish and secure collection points for evacuees
- Provide secure air and/or land transportation for evacuees from collection points to points of departure
- Deploy sufficient forces to overcome all plausible resistance
- And limit damage to relations with existing—and perhaps surviving— government and avoid prematurely prejudicing U.K. relations with a future Egyptian leadership

Constraints

The evacuees are widely dispersed in heavily populated areas. Strict rules of engagement (fire only when directly threatened) must be maintained to avoid unnecessary conflict with Egyptian forces and minimize casualties to Egyptian civilians. The Egyptian government's operations against the rebels present major uncertainties in determining the friendly or hostile status of host-nation forces at the lowest levels (individual aircraft, ships, air-defense batteries, and ground-force units from platoon size up). The aerial and seaports of embarkation are not secured. Basing access is available only in Israel and Turkey.

Task Execution

In order to execute this mission, U.K. military leaders decide that deploying AI combatants for the rescue mission is the best option. Given the large success of Levia Watch and Strike, the MoD has acquired a new version of this ground unmanned vehicle: the Levia Rescue (Levia R). Levia R is a multifunctional tracked robotic system that can be armed with AR15s, such as M4a1, CQBR, and Colt Commando guns to be used for defense (these are not offensive strikers). The unmanned system is a 7.5ft-wide, 4ft-tall tank-like robot with speed reaching 20mph. The tank is able to carry various heavy payloads, including up to two adult humans. Making technical progress from the generation 1 Levia Sentry and Strike, there no longer is a human in the loop for the Levia R; human military personnel are on the loop and so, once the AI is deployed, humans have limited capabilities in controlling its actions.

During the rescue mission, a Levia R agent misassesses a human-robot interaction as aggressive. The human action is perceived as threatening and so the agent uses its defense technology to shoot. The human is killed.

It is soon discovered the human is a non-combatant -- a civilian who was not engaging in any threatening acts towards the unmanned system (the Levia R's 'body' camera footage was reviewed by human military personnel and the act of the civilian was deemed to be non threatening). There was no external intervention with the AI system given that humans are sitting on the loop. The misassessment is being viewed as a system failure, though the reasons of why the system failed are uncertain.

Public Knowledge

During the altercation, another civilian recorded the interaction on their mobile device and has posted it on social media. The video is shaky and the altercation goes in and out of frame and so the exact actions of the human and robot are slightly uncertain. The final shot is captured.

Assumptions

1. There is no AGI (narrow AI only -- applications for narrowly defined tasks).
2. There are no existing international or domestic bans on LAWS.
3. The movement for private companies to refuse engagement with LAWS has failed to spread effectively - companies like Amazon, Microsoft, etc are still engaged.
4. The state of technology (SoT) has passed domestic technology readiness standards.
5. There has been relatively widespread integration of LAWS in various missions (targeting, search and rescue, and more).