Written evidence submitted to UK Parliament Artificial Intelligence in Weapons Systems Committee Select Committee Inquiry, April 2023

Breeze, Emma

UK Parliament Artificial Intelligence in Weapons Systems Committee
Select Committee Inquiry, April 2023

Written evidence submitted by Dr Emma J Breeze (formerly Marchant) at the University of Birmingham.

Introduction

I am an Assistant Professor in International Criminal Law at the University of Birmingham. My research focusses on the use of information during armed conflict and the impact of new technologies on International Humanitarian Law (IHL). My doctoral thesis (2020) and current research directly relate to autonomous weapons systems (AWS) specifically related to compliance with the precautionary principle of International Humanitarian Law (IHL). I have investigated the precautionary principle of IHL to develop an intelligence standard for targeting during armed conflict, which is essential for AWS compliance with IHL.

Executive Summary

This evidence responds to Question 4 of the Call for Evidence only. It outlines the legal accountability of autonomous weapons systems (AWS) under the precautionary principle of International Humanitarian Law (IHL). The precautionary principle brings effect to the overarching legal obligations of distinction and proportionality, and without consideration of its mandate it will be difficult for AWS to comply with IHL.

This evidence outlines how the precautionary principle regulates targeting, how this is understood and why this is problematic for AI-driven systems. Four recommendations are then provided to assist in the development of these systems to ensure compliance with IHL.

Committee question 4:

Is existing International Humanitarian Law (IHL) sufficient to ensure any AWS act safely and appropriately? What oversight or accountability measures are necessary to ensure compliance with IHL? If IHL is insufficient, what other mechanisms should be introduced to regulate AWS?

1. In principle IHL governs the legal compliance of the operation of any AI-enabled technology that is deployed during armed conflict. However, it must be recalled that IHL is limited to situations that amount to an armed conflict and thus the wider deployment of AI, notably to law enforcement or other organisations outside of a time of armed conflict, would need to be considered under other legal principles, such as International Human Rights law.

2. In general terms AWS are not in themselves unlawful under IHL but they can operate in a manner that would be unlawful. Therefore, the development of the technology and deployment of such is not problematic for IHL (presuming Article 36 reviews are conducted), it is in how they are operated that creates difficulties for compliance with IHL. Therefore, in the Autonomy Spectrum Framework presented by the Defence Artificial Intelligence Strategy IHL is sufficient to govern the use of AWS that are 1. Human Operated or 2. Operator Assisted. Indeed, these types of weapons systems have been in use for several decades, such as the US Patriot Missiles and

Phalanx Weapons Systems. These both have autonomous features but are controlled by a human operator, either in a decision or authorisation capacity. However, as the level of autonomy advances through 3. Task Autonomy, 4. Conditional Autonomy and finally 5. Highly Autonomous, the challenges for compliance with IHL increase.

3. In the foreword to the Defence Artificial Intelligence Strategy Policy Paper of June 2022, there are three distinct 'imaginings'; a Solider using AI within a command-and-control network, a logistics supply undertaken with and through AI, and finally the use of AI for directed energy weapons in defence. Primarily this evidence will discuss the first scenario relating to information and decision-making during targeting, with some consideration of the use of AWS for defensive applications. This evidence will show that IHL may be challenged when AI is used beyond that of automated lethality.

The Precautionary Principle of IHL

4. As detailed within the related Defence Policy Papers and the UK MOD Joint Service Manual of the Law of Armed Conflict, the basic principles of IHL are those of Military Necessity, Humanity, Distinction and Proportionality. To meet these overarching principles further obligations are established by IHL, notably regarding the gathering and dissemination of information under the precautionary principle.

5. This principle is stated by the UK MOD at 5.32 in the Joint Service Manual, repeating the detail from the Geneva Conventions at Article 57 of Additional Protocol I 1977. It is customary international law and requires states to "do everything feasible to verify that the objectives to be attacked are neither civilian nor civilian objects…" The obligation also brings detail to collateral damage and the proportionality assessment saying that states must "refrain from deciding to launch any attack which may be expected to cause incidental loss of civilian life… which would be excessive in relation to the concrete and direct military advantage anticipated." Thus, the precautionary principle is intrinsic to the targeting process and provides the basis for compliance with IHL. As such, any AWS would have to meet the precautionary principle and be able to demonstrate that it took 'all feasible precautions' prior to and during an attack.

6. The standard required by the 'all feasible precautions' obligation is not absolute and can be understood as those precautions which are practically possible in the circumstances prevailing at the time. It is a proactive but contextual standard, with no absolute obligation on the intelligence gathering system to produce accurate information. However, it does require those who plan or conduct attacks to do everything feasible to verify targets are military objectives and this could include gaining more information. It should be noted that in cases of doubt IHL presumes a civilian status, but the United States DoD manual does not recognise this as customary law. This shows how regular allies may interpret aspects of IHL in a disparate manner.

7. Nonetheless, states have an obligation to ensure that their targeting protocols and wider military operations take all feasible precautions prior to launching an attack, as well as during that attack. Both the strength and weakness of the precautionary principle is its inherent imprecision. It is not a 'bright line' rule and is qualified by that which is practically possible considering the prevailing conditions, which can vary considerably during a period of armed conflict. Due to these qualifications the 'feasible precautions' obligation can adapt to modern demands, despite considerable technological development since codification. However, it now arguably requires a higher standard than previously, with states required to use all information available to them to make targeting decisions, prior to and during an attack.

8. This can be rather complicated in practice, and it is increasingly difficult to establish what standard of intelligence is required prior to launching an attack. For example, in 2014 Israeli Defence Forces were criticised by the UN for conducting an attack that relied upon UAV surveillance footage. The UN commented to the BBC that "the resolution of the video is so poor compared with proper satellite imagery that you cannot see some of the trees in the compound, let alone people." The difficulty presented here is that the UN have compared the footage presented by the Israeli Defence Force to 'proper satellite footage' this implies an expectation of a higher quality of information is required. However, as stated, IHL does not provide a quality or quantity standard for intelligence merely that which is available in the prevailing circumstances.

9. This example demonstrates the difficulties presented by the sliding scale of precautions as required under IHL. It also highlights the challenge that precautions are not applied identically by different states, for example the US since the late-1980s have used a 'positive identification' (PID) standard which uses the phrase 'reasonable certainty'. This shows that for interoperability it is critical for UK AI systems to comply with a clear 'intelligence standard' that can be adapted as needed when deployed in a coalition.

10. For autonomous systems to be able to comply with IHL principles it is crucial that we can provide a version of the legal obligations that can be adopted by these systems. This also needs to be accepted and understood by developers and partners as the reliability of these systems and the persons responsible for their development, production and operation need to be clear what these obligations require.

11. Furthermore, although it is presently accepted that autonomous systems will remain compliant with IHL by having a 'man-in-the-loop' or with 'meaningful human control', my concern is that this 'man in the loop' is being provided with intelligence information that has been analysed and filtered by algorithm, thus potentially distorting situational awareness. Furthermore, I believe that the problem of deception within the intelligence cycle is a potentially significant problem that could easily mislead machines and provide information to operators that is incorrect, whether by nefarious means or purely that which has been treated as true by the system.

AI in Data Analysis and Decision Making

12. This becomes particularly relevant when considering the Defence Artificial Intelligence Strategy at 5.2.3 which discusses the role of 'Automation in Data Analysis'. It is recognised that AI can be beneficial in the gathering and processing of large quantities of data to enable rapid 'integration, exploitation, and production of intelligence'. The danger here is in the conflation of information, the infiltration by malicious information and the lack of transparency in the data that is used in reaching decisions to use lethal force against targets. A good example of this comes from Takhar, Afghanistan in 2010. The target of this strike was an individual understood to be the Taliban's shadow governor for the Takhar region, known as Mohammed Amin. Amin was placed on the Joint Prioritised Effects List by ISAF's Joint Special Operations Command (JSOC). The intelligence operation that led to him being placed on the list was as a result of mapping a cluster of cell phones related to the Taliban and Islamic Movement of Uzbekistan and their monitoring. The analysts came to believe that one of the SIM cards they were monitoring had been passed to Amin and he had started to use the name Zabet Amanullah as an alias or 'nom de guerre'.

13. The individual targeted in the convoy was a man known as Zabet Amanullah, and, evidently, he was carrying the cell phone that was being tracked by the US. The problem is that the Zabet Amanullah who was travelling in the group was travelling as part of a parliamentary election

convoy. The district governor of Rostaq, Malim Hussain, confirmed that the convoy belonged to the candidate Mr Khorasani who was travelling in the area. Hussain was reported as saying that as a result of the attack "ten people were killed, including a local commander called Amanullah, a former member of the Mujahideen who was not a member of the Taliban." The crucial fact then is whether the agent travelling in the convoy known as Zabet Amanullah, who was using the phone tracked by the US and a former member of the Mujahideen, was, in fact, Mohammed Amin.

14. To date ISAF remain certain that they targeted the right man, but all those who have investigated the incident, and knew those involved, have concluded that somewhere along the line the identities of Amanullah and Amin became conflated. This incident highlights the weaknesses in automated analyses that rely upon compartmentalised information, as well as a lack of understanding by operators in how that information and decision has been reached. The individual here had led a high-profile public life which could have been confirmed with a simple internet search. However, once Amanullah was placed on the list and his SIM was geo-located he was effectively in the crosshairs. Thus, a reliance on AI to generate, develop and produce information needs to be very carefully managed to ensure transparency remains.

15. The linkage between the data-driven information and the addition to a 'kill list' effectively undermines the precautionary principle of IHL. In the case of Takhar, once Amanullah was placed on the list the operators considered precautions to have been taken in identifying him as a military target, and thus the principle of distinction was met. However, that had, in effect, been covered by a computer network that was solely relied upon to affirm the nature of the target. Given the rise of information warfare and an increasing use of open-source and social media intelligence (OSINT and SOCMINT) the potential for mistakes within data, and the subsequent targeting errors also increase.

16. Whilst these mistakes may happen irrespective of AI, the wide scale and scope of AI could increase their likelihood, and place both civilians and military personnel in danger from conflated, misleading, or manipulated information. To return to our solider being provided information for control-and-command it is important to question who is in control of that operation. Whilst the soldier is ultimately making the decisions, and for IHL purposes can be held individually criminally responsible for any serious or grave breaches of IHL, there is substantial evidence that human-machine interactions are far from perfect. Early examples of the over-reliance and faith in machines can be shown from the mistake of shooting down civilian airliner Iran Air 655 by the USS Vincennes' in 1988 and the loss of an RAF Tornado due to friendly fire in 2003.

17. In both examples semi-autonomous weapons systems were acting in a defensive capacity, much like the imagining of the Defence Strategy, and misidentified the aircrafts as an enemy rather than civilian or friendly, a misfunction of the IFF system. In both cases the human commanders failed to interrogate the systems and relied upon the data they were provided resulting in fatalities. Thus, whilst the commanders were obligated to meet the precautionary principle of IHL they did not question the abilities of the systems to do so. It is suggested that to enable defensive AWS to comply with IHL they can only be used in a capacity which 'fails-safe' or indeed requires an active decision by a human. However, without appreciating and mitigating the limits of human abilities in these situations the effects will remain.

Other Measures

18. Consideration should be given to the Rules of Engagement (ROE) developed for use by and with AWS. These are the realisation of IHL operationally by military forces and can be more restrictive than IHL. These could be used to ensure AWS and Automated Decision Making with AI comply with IHL, for example by restricting the use of AI in populated areas, or during law enforcement type operations. However, without a clear understanding of the requirements of 'all feasible precautions' it would be difficult to define parameters that would be practical whilst balancing military necessity and the principles of humanity.

AI and Precautions

19. In principle AI has the ability and possibility to improve situational awareness, it has the potential to reduce civilian casualties and mitigate mistakes. An example of this could be the Gredlica Bridge incident from Kosovo, in which an aircrew mistakenly bombed a passenger train as it was crossing the bridge. In the footage it is shown that the first missile hits the bridge, but the second had already been sent by the time the passenger train appears. In these minutes an automated system could perhaps have saved the train. That said, for the system to have intervened it would need to understand the IHL obligation to cancel or suspend an attack and have the ability and authority to intervene.

20. The overarching claim of the 'new technological gospel' has been that computers, advanced sensor platforms, satellites and the persistent eye-in-the-sky will dispel the 'fog of war' by eliminating friction and uncertainty presented by adverse terrain, climate, morale, equipment failure, and other factors. To date technology has not removed the 'fog of war' despite its best efforts. The rise in intelligence, surveillance, and reconnaissance (ISR) platforms has given a significant advantage to states with highly advanced assets but it has also increased their responsibility for compliance with the sliding scale of all feasible precautions required by IHL. In the same way precision munitions raised expectations of accuracy AI decision-making will increase the demand and expectation of accuracy in target identification. When this fails it will damage trust in the systems by military and civilian stakeholders, it will cost human lives and ultimately it will result in a loss of legitimacy.

21. Recommendations

This evidence demonstrates that for AWS to comply with IHL they must be able to understand and apply the 'sliding scale' of the precautionary principle. There is no 'bright line' rule on the quality and quantity of intelligence required for positive verification, and different states have different understandings of their obligations. Additionally, this evidence indicates how AWS affect lethal outcomes through data analysis and defensive actions. To meet IHL obligations, and to assure legitimacy and trust in operations, UK AWS need to consider not only the final lethal act but the preceding actions that lead to the decision to target objectives.

a. Establish understanding of feasible precautions under IHL for targeting decision making as a legally obligated 'intelligence standard'. Ideally this would be an internationally agreed standard, but a UK understanding would be substantial progress and increase interoperability and international legitimacy.

b. Consider and mitigate the links of any AI system to lethal outcomes, including the listing of individuals on target lists. Review practices related to target identification and the dissemination of such, including where UK intelligence may be passed to partners.

c. Produce guidance and procedures for fail-safes for AI-controlled defensive systems to prevent significant loss of life due to algorithmic error, or infiltration by rogue data. An option could include sacrificial directives to AWS, where in doubt the AWS risks its own destruction rather than risk human life.

d. Promote legitimacy and trust through transparency in usage and procedures, with full investigations established to ascertain best practice and report errors. These investigations will support the development of technology and promote respect for IHL by partner states, as well as promote trust and legitimacy domestically.