Written Contribution by the AutoNorms Project

Submitted to the Chair of the Group of Governmental Experts (GGE) on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (LAWS)

8 September 2021

The European Research Council (ERC)-funded AutoNorms project[1] is based at the Centre for War Studies, University of Southern Denmark.

Response to the Chair's Guiding Question

How would the analysis of existing weapons systems help elaborate on the range of factors that should be considered in determining the quality and extent of human-machine interaction/human control/human judgment?

Weaponised artificial intelligence (AI) raises significant ethical, legal, and political questions – many of which are considered to be negative. Much of the current debate on the development of Lethal Autonomous Weapons Systems (LAWS) within and beyond the GGE frames these issues as concerns for the future which require a *preemptive* framework to manage and regulate. We believe that this approach is misguided. It distances the debate on LAWS from weapons systems that already have automated and autonomous features in their targeting functions. These include guided missiles, air defence systems, active protection systems, counter-drone systems, and loitering munitions.

Some of these types of systems are far from new: close-in weapons systems, for example, have integrated automated features since the 1970s and the level of automation has steadily increased thereafter. In some instances, the use of these

[1] The Principal Investigator of the AutoNorms research project is Dr Ingvild Bode, Associate Professor at the Centre for War Studies, University of Southern Denmark. The project is funded by the European Research Council (grant agreement no. 852123). For more information about the AutoNorms project and research updates, please visit our website: www.autonorms.eu

technologies has meant that human control over the use of force has become *meaningless*, a phrase which we use to capture two connected dynamics. First, the inability of human agents to exercise deliberative control over certain weapon systems because of the speeds at which these systems operate, the complexity of the tasks they perform, and the demands human agents are placed under (i.e. human control over the use of force lacks significance). Second, as the cumulative effect that the incremental integration of more autonomous and automation features has had on reducing the range and substance of meaningful human control in specific targeting decisions (i.e. human control has come to mean less over time).

Because of this, we encourage states parties to the UN Convention on Certain Conventional Weapons (CCW) to think about the integration of autonomy and automation into the targeting functions of weapons systems along a much longer trajectory. Trends which on first glace may appear 'new' have, in fact, rather a considerable history deserving of scrutiny. This history is particularly important within the context of the GGE's discussions on LAWS because the integration of automated and autonomous features into weapons systems has already shaped understandings of the appropriate quality of human control in specific targeting decisions.

The AutoNorms project argues that examining existing systems provides a crucial entry point for understanding the changing nature of human-machine interaction and the challenges that autonomous features in targeting pose for retaining meaningful human control over the use of force. Examining such challenges through the detailed study of different types of weapons systems is a key objective of the AutoNorms project. We have completed a study of air defence systems[2] based on two sets of empirical data: a data catalogue of automated and autonomous features in 28 air defence systems;[3] and a close examination of human-machine interaction in four different air defence systems involved in high-profile failures that brought down civilian and military aircraft in friendly fire incidents.

Our research shows that the role of human operators has been fundamentally changed through integrating automated and autonomous functions into air defence systems.

[2] Ingvild Bode and Tom Watts, "Meaning-less human control: Lessons from air defence systems for lethal autonomous weapons", Oxford & Odense: Drone Wars UK & Center for War Studies, February 2021, https://dronewars.net/wp-content/uploads/2021/02/DW-Control-WEB.pdf

[3] Tom Watts and Ingvild Bode, "Autonomy and Automation in Air Defence Systems Catalogue," February 2021, DOI: 10.5281/zenodo.4485695.

The major qualitative change is that the role of the human operator has been minimised while, simultaneously, becoming increasingly complex. Our research demonstrates that designing, training personnel for, and operating air defence systems with automated and autonomous features in targeting have contributed to an emerging norm that diminishes the quality of human control over specific targeting decisions. The human operators' roles in air defence systems have changed from active controllers to passive supervisors. This has meant that they have lost both situational awareness and a functional understanding of how algorithmic systems make targeting decisions. While human operators often formally retain the final decision, in practice the decision that is made based on information from highly complex systems in fast evolving situations is often *meaningless*. This diminished role of human control has been gradually normalised over time.

This emerging norm has been shaped in a silent process for how states have designed, trained personnel for, and operated air defence systems with automated and autonomous features. This process precedes the international debate at the CCW by decades and continues to run parallel to it. The debate on LAWS has yet to scrutinise this emerging norm. Currently, if air defence systems or other existing weapon systems with autonomous or automated features are mentioned at all, they are not considered to pose challenges to human control: states can limit where, how, and when they deploy air defence systems by setting their parameters of use and controls on the environment. Further, air defence systems have human operators in-the-loop or on-the-loop in specific use of force decisions. But our research demonstrates that being in/on the loop does not guarantee that human operators can exercise meaningful human control due to the complexities of human-machine interaction. Not acknowledging these processes undercuts potential international efforts to regulate LAWS through codifying an appropriate quality of human-machine interaction.

<u>Recommendations</u>

To help facilitate critical reflection, the AutoNorms project supports new international law on autonomous weapons systems based on meaningful human control as a central, positive obligation. To help ensure that such legal safeguards ensure *meaningful* rather than *meaningless* human control over the use of force, we make four recommendations for stakeholders involved in the GGE debate . These begin from the premise that positively codifying an "operationalized" version of meaningful human control is the most promising avenue for creating a regulatory framework on LAWS' development and usage.

- Practices of human-machine interaction associated with existing weapons systems that have automated and autonomous features in targeting should be openly scrutinised.
- The study of existing weapons systems can provide practical insights into the existing and future challenges to human-machine interaction created by autonomy and automation that, if not explicitly addressed, may shape silent understandings of appropriateness regarding these technologies. We support calls by stakeholders such as the ICRC and SIPRI for the detailed study of existing autonomous weapon systems including, but not restricted to, loitering munitions.
- The study of air defence systems highlights that while all three components of meaningful human control (technological, situational, and human-machine interaction) are important, control through human-machine interaction is a decisive element in ensuring that human control remains meaningful. This is not least because human-machine interaction highlights meaningful human control at the specific point of using a weapon system, rather than the exercise of human control at earlier stages, such as during research and development.
- Control through human-machine interaction should be integral to any codification of meaningful human control. AutoNorms identifies three prerequisite conditions needed for human agents to exercise meaningful human control:
  (1) a functional understanding of how the targeting system operates and makes targeting decisions including its known weaknesses (e.g. track classification issues);
  (2) sufficient situational understanding;
  (3) the capacity to scrutinise machine targeting decision-making rather than over-trusting the system.

These three prerequisite conditions (functional understanding, situational understanding and the capacity to scrutinise machine targeting decision-making) of ensuring meaningful human control in specific targeting situations set hard boundaries for AWS development that should be codified in international law. In our assessment, they represent a technological Rubicon that should not be crossed as going beyond these limits risks making human control *meaningless*.