Statement of the UN Institute for Disarmament Research

on Lethal Autonomous Weapon Systems
at the Fifth Review Conference of the CCW
13 December 2016
Delivered by Kerstin Vignard, Deputy to the Director

Mr Chairman, distinguished colleagues,

In support of next steps on the issue of LAWS, I would like to share with you a few observations from UNIDIR's latest report, "Safety, unintentional risk and accidents in the weaponization of increasingly autonomous technologies".

Much of the LAWS discussion in the CCW has—so far—focused on whether their introduction would create a legal 'responsibility gap' in targeting and individual attacks. Apportioning responsibility for this is a critical concern. But is it enough?

No complex, tightly coupled technology is fail-safe, even when carefully designed and operated. Nevertheless, as detailed in the paper, we found that AWS may pose novel, unintended forms of hazard to human life that typical approaches to ensuring responsibility may not effectively prevent. In systems using machine learning techniques, in particular, the *observability* of the internal processes may be low—and timely human *directability* in the case of errors rather limited. This can add up to *unpredictability*.

Experts themselves have begun to call for more attention to understanding and preventing machine learning-related accidents.

At present, because many of the applications of autonomous systems are in their infancy, the consequences of such accidents are probably limited. However, this is likely to change, maybe sooner than we may expect, especially if there is further convergence between the development of autonomous systems based upon machine learning and interest in weaponizing them.

Where machine learning is being used in civil systems such as autonomous cars, a hybrid of machine learning, handcrafted rule sets and 'fail-safes' are used. Even then, formally verifying the behaviour of machine learning sub-systems is a risk mitigation challenge. For instance, Google has put its autonomous vehicles through more than three million kilometres of road testing in real driving conditions, and has said that considerably more testing will be needed before it puts such vehicles into production. It is hard to envisage what an equivalent level of testing and fail-safes would look like for AWS.

Nevertheless, the outcomes of AWS failures could include fratricide, civilian casualties, or unintended escalation in a crisis as machines pursue emergent yet inexplicable goals such as area denial. This could also have strategic consequences.

As explained in the paper, having **humans in or on the loop** is not a solution in itself to safety-related risks. In many cases, the human operator is expected to understand how the complex system works as well as anticipate its actions. Ultimately, this can increase their responsibility. Beside the practical hazard this kind of practice can create, it is also raises the question of whether humans should be placed in situations in which they are held

accountable for the behaviour of armed learning systems they might not be able to understand or fully control.

Governments are responsible for creating sound policies on which uses of autonomy are legitimate in weapon systems. Advances in relevant technologies and mounting public concern are adding to the pressure to do so. We encourage the High Contracting Parties to consider all facets of the issue of responsibility in your future discussions—including the consequences that arise from accidents and unintended risks in systems dependent on machine learning for critical functions. UNIDIR's latest paper serves as one resource. This, and the project's earlier papers and audio files from public events, are available on our website. I would like to thank Canada, Germany, Ireland, and the Netherlands for their investment in this work at UNIDIR.

UNIDIR, for its part, will continue to support your work. In 2017, in anticipation of the establishment of a GGE, our work will have a particular focus on definitions and characteristics of AWS, mapping friction points in near term tech development, as well as the impact of biases and values on learning systems.

Thank you, Mr Chairman.