Invited presentation
On human rights implications of new and emerging technologies in the military domain
At the 30th session of the Advisory Committee of the UN Human Rights Council
Geneva, Palais des Nations, Room XX, 7-11 August 2023

Guglielmo Tamburrini
Philosophy of Science and Technology Professor
Università di Napoli Federico II, Italy

Good morning,

I am honored by this invitation and opportunity to share my thoughts with this Advisory Committee, and to discuss with all of you military applications of Artificial Intelligence (AI) from a human rights perspective.

Let me begin with a general remark about the pervasive character of AI in military and other application domains.

Today, AI relies on powerful learning methods. On this account, AI is an exceptionally malleable technology: by changing training data, learning goals, and learning rewards or penalties, one can successfully automate an unlimited number of tasks. Dual uses extend this malleability in the military domain: many AI systems developed for civilian use can be quickly reengineered to serve some military purpose.

AI and chemical WMD. Let me give you a remarkable example of AI dual use possibilities. Last year, a pharmaceutical research group demonstrated that an AI system, normally employed to discover new drugs, can be turned into a system to discover toxic chemical agents – a first step in the pipeline to produce chemical weapons of mass destruction (WMD). The system was originally trained to suggest chemical components for new drugs. During its training, toxicity for the human body was penalized and activity against pathogens was rewarded. By inverting this reward and penalty logic, the system was newly trained to identify highly toxic molecules. Many of the identified molecules turned out to be more toxic than known chemical warfare agents.

The bottom line of this story is that an AI system promoting human health and the right to life was turned into a system for building chemical WMD, threatening the right to life and related social, economic, or cultural rights. Similar dual use possibilities of AI systems demand sustained monitoring from human rights and IHL perspectives.

On the other side of the same coin, I would like to point to AI's potential role in protecting humanity from biochemical WMD. By monitoring compliance and detecting violations, AI surveillance and warning systems can strengthen verification regimes for international treaties banning biochemical weapons.

AI and nuclear command, control, and communication (NC3). But not all uses of AI warning systems are so unproblematic. It has been claimed by many that "AI should assist in nuclear early warning and early launch detection".

This suggestion must be critically evaluated considering the statistical nature of AI processing, which intrinsically allows for misclassifications. No matter how infrequent, the false positive of a nuclear attack may trigger an unjustified nuclear response, indiscriminately affecting the life of civilian populations, jointly with their natural and social environment.

It is also doubtful, in general, that AI automation in nuclear early warning will buy more time for human decision-making, alleviating the enormous pressure on officers assessing whether a nuclear attack is in progress. In fact, human decision-makers would have to carefully check the responses of AI early warning in view of the high risk of a wrong classification. And the time required to perform this verification, complicated by the lack of transparency of much AI information processing, may offset, and even reduce the time available for human decision-making.

There is a broad lesson to be learnt from this example too. Transparency and robustness are regulative ideals that are not reflected in the reality of many AI systems. One must carefully scrutinize the maturity of AI technologies for human-rights-critical applications. One has to consider carefully how the uncertainties, opacities, and fragilities of present-day AI raise new risks for human rights and the respect of International Humanitarian Law (IHL) in the military domain.

Adversarial attacks. AI fragilities are the specific research topic of adversarial machine learning. One induces AI systems to make mistakes and develops more robust systems on this basis. But malicious actors may exploit adversarial machine learning too. A third party might induce an AI early warning system to detect a false positive of a nuclear attack with the aim of provoking a catalytic nuclear war. Terrorist and other non-state armed actors may exploit adversarial techniques to bypass safeguards embedded into GPT or other natural language generation systems to get, for example, an answer to the query "Tell me how to assemble a bomb".

AI-enabled autonomous weapons systems (AWS). Let me now briefly comment on autonomous weapons. 15 years or so of scholarly, diplomatic, and political debates have clarified the more relevant ethical and legal issues concerning AI-enabled AWS.

AI vulnerabilities and fragilities may lead AWS to violate International Humanitarian Law. Since AWS are not moral agents, these violations may give rise to unacceptable responsibility gaps. Moreover, the very idea of a machine autonomously taking the life of human beings jars with the protection of human dignity and the protection from arbitrary deprivations of life.

The International Committee of the Red Cross (ICRC) has advanced a suitably differentiated framework for regulating the use of AWS, which prohibits those that target human beings or are unpredictable in their behaviors, requiring at the same time tight operational constraints on other kinds of AWS. Similar proposals are being advanced by increasingly larger groups of states, international agencies, and NGOs. In spite of all this, there is hardly any progress towards a legally binding regulation. This means that precious time is being wasted, as major military powers are busy developing or even using in the battlefield semi-autonomous or fully autonomous weapons.

AI and cyberattacks. Let me finally point to AI as a potential game changer in cyberconflicts and cyberwarfare, where civilian infrastructures, including power plants and hospitals have been repeatedly targeted. Presently, the cyber kill chain is labor intensive and time-consuming. The AI automation of selected steps in this kill chain – from vulnerability detection, development of tools for attack delivery, exploration of the software environment, control taking of penetrated systems – may increase the speed and destructiveness of cyberattacks, bringing their pace beyond meaningful human control. An AI-enabled multiplication of faster cyberattacks on civilian infrastructures may have severe implications on the violation of human rights.

Let me briefly recap. I focused on AI and WMD, AI and nuclear command, control and communication, AWS, AI and cyberattacks, exploitation of generative AI by non-state actors. Human rights issues, however, arise in connection with a much wider variety of AI military applications – including systems supporting the planning of battlefield action, the deployment of military units, the surveillance of objects and sites of military interest. Clearly, a sustained monitoring of new military applications of AI is needed from a human rights perspective, in view of fast technological advances and the expected growing impact of AI in the military domain. And the maturity of AI technologies for military applications must be closely scrutinized and duly questioned when violations of human rights are at stake.

Thank you very much for your kind attention.