Risks

As AI-robots are becoming more prevalent in warfare, and are undergoing rapid development with greater speed, operational scale, autonomy, and lethality, it is imperative to critically assess the limitations of current AI-robots that induce risk on the battlefield.

Recent studies have demonstrated that neural network-based AI computer vision algorithms can be manipulated to alter object classification by modifying just a small number of pixels in the image (Elsayed et al., 2018). In relation to the battlefield, adversaries could exploit such vulnerability of AI systems to modify classification outcomes. This manipulation could lead to unintended actions by the AI-robots, posing a potential risk during military operations.

For trust to be established in an AI-empowered combat robot, it is crucial that the robot consistently performs tasks effectively and in a manner that can be understood by the operator, and essentially remains impervious to manipulation by adversaries. If cyberattacks enable adversaries to gain control over the robot, it would undermine the operator's trust in the robot's ability to fulfill missions as instructed, thereby weakening its operational effectiveness.

In 2018, Russia made a claim that the United States had taken command of 13 UAVs and allegedly intended to deploy them in an attack on a Russian military base in Syria (The Associated Press 2018). Ensuring cyber resilience is crucial for the effective deployment of AI-robots on the battlefield, as the presence of vulnerabilities that could allow AI-robots to be redirected or manipulated from their intended purpose would pose an unacceptable risk to both friendly forces and civilians.

Another major risk is that current AI lacks the ability to grasp the contextual aspects of actions, behaviours, and relationships (von Braun et al., 2021). This intricate web of experience necessary for robust reasoning and decision-making remains beyond the reach of current AI systems. As stated by Robitzski (2018), the development of an AI with human-level reasoning is a long-term objective that remains uncertain in terms of the timeframe, with estimates ranging from 10 years to an indeterminate period.

With the present inability of AI to reason in such high-stakes settings, such limitations on domain keep it from fully understanding the repercussions of its actions. Greater autonomy without deep understanding may potentially lead to catastrophic escalations and further boost the speed of warfare.

Further, while computer vision algorithms have made remarkable strides in identifying individuals and objects in video data, their focus on object detection and classification falls short of comprehending the three-dimensional world (Swett et al., 2021). Figuring out the world in three dimensions requires more than just detecting and classifying objects. One of the main technical challenges in perception involves consistently tracking humans and objects in diverse backgrounds and lighting conditions, accurately identifying, and tracking individuals across different contexts, recognizing behaviours, and predicting intentions or outcomes (Swett et al., 2021). These complexities highlight the ongoing work needed to overcome these technical hurdles and achieve a deeper understanding of the environment through AI systems.

At present, AI robotic systems lack "a sophisticated multi-sensory perceptual system that can create rich and detailed internal representations of the state of the world and its dynamics, in real-time." (von Braun et al., 2021). This limitation in perception gives rise to several risks, such as the inability to consistently differentiate between friend and foe, accurately identify the relationships between various elements in a scene, and fully comprehend the implications of the actions, individuals, and objects within a given context.

This raises questions on the chain of command as well. In the event of improper conduct or unauthorized harm caused by an AI robot, whether due to errors or intentional actions, determining who should be held accountable and subject to punishment becomes a complex issue. This presents a significant obstacle as no human individual can properly be held accountable for the outcomes they generate (Asaro, 2013). As a result, the existing legal frameworks fall short of addressing the accountability and liability aspects associated with AI robots, necessitating the development of comprehensive and coherent guidelines to navigate this intricate landscape.

Overcoming these limitations and addressing technical challenges are essential to ensure the safe and effective deployment of AI-robots in military operations. As AI-enabled autonomy continues to evolve, it is crucial to strike a balance between harnessing its potential and mitigating the risks it presents to maintain the safety and security of both friendly forces and civilians. Understanding the constraints of current AI-robots and their future direction enables policymakers to make informed decisions regarding the appropriate uses and potential risks of misuse.

References:

Asaro, Peter (2013). *On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making.* *https://doi.org/10.1017/S1816383112000768*

Elsayed, G., Shankar (2018). *Adversarial Examples that Fool Both Computer Vision and Time-Limited Humans*. http://papers.nips.cc/paper/7647-adversarial-examples- that-fool-both-computer-vision-and-time-limited-humans.pdf.

Robitzski, D. (2018). *When Will We Have Artificial Intelligence As Smart as a Human?* *https://futurism.com/human-level-%20artificial-intelligence-agi*.

Swett, Bruce A., E. N. Hahn, A. J. Llorens (2021). *Designing Robots for the Battlefield: State of the Art* https://library.oapen.org/bitstream/handle/20.500.12657/47279/9783030541736.pdf?sequence=1#page=131

The Associated Press. (2018). Russia claims US aircraft took control of drones in attempted attack on its Syrian base. https://www.militarytimes.com/flashpoints/2018/10/25/russia-claims-us-aircraft-took-control-of-drones-in-attempted-attack-on-syrian-base/.

Von Braun, Joachim (2021). *Robotics, AI, and Humanity: Science, Ethics, and Policy*. Springer. https://library.oapen.org/bitstream/handle/20.500.12657/47279/9783030541736.pdf?sequence=1#page=131