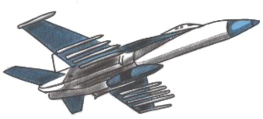
iPRAW
International Panel on the Regulation of Autonomous Weapons

Scenario for Workshop I – Concepts of Human Control

October 27th, 2020

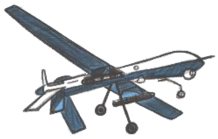
SCENARIO

The Mission



Arcadia and **Utopia** are engaged in an international armed conflict. Arcadia's military commander in chief decides to **attack one of Utopia's most important military bases** in Utopia's south. However, the military base is protected by a modern integrated multi-layer **air defense system** to detect, track and engage hostile aircraft. Utopian forces are using both mobile and stationary systems, which are arranged to build an overlapping air defense area consisting of separate clusters. Whereas the assets proximate to the operating **base are not collocated with non-military targets**, the distributed, mobile air defense units in the **advanced posts are partly deployed to urban areas**, where civilians and non-military objects could be harmed. However, the neutralization of the air defense system as a whole is key in order to gain access to the operating base. For this reason, the military commander in chief tasks the Arcadian air component to engage the Utopian air defense system. One deployed **fighter jet is accompanied by five armed UAVs.** The UAVs are equipped with **jamming capabilities** using radio frequency to interfere with the air defense system's radar, missiles to neutralize the targets, various sensors like a high-resolution video camera, radar, an infrared camera as well as software able to interpret video footage within seconds. **The software is able to identify the air defense system and can differentiate between objects and persons.** Prior to acquiring the weapon system Arcadia has undertaken a careful weapons review and the military personnel was trained in order to be able to understand and operate the system.

The Technology



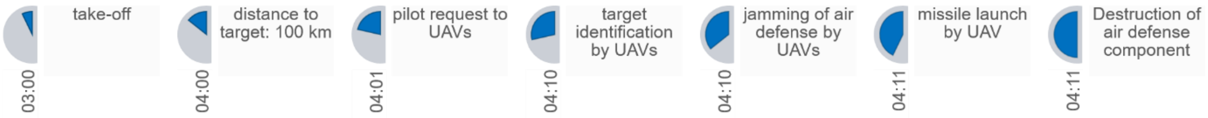
Based on data from the UAVs' cameras, the UAVs' image recognition program is requested to identify the air defense system as the main target and to ascertain whether people are present in the vicinity. According to the manufacturer of the UAV, the recognition software has a **success rate of 89%** when it comes to the differentiation between objects and persons. In 11% of the cases the software does not identify any persons to be present although the opposite is true. In case the program does not detect persons, the UAV, being the sensor and shooter, will launch an attack by deploying the missiles installed on it. **No further human intervention is necessary in this case and the fighter pilot exclusively relies on the software's assessment.** If the UAV detects the presence of humans, it has to ask the pilot for further instructions by sending footage that was taken by the video as well as thermal imaging cameras to the pilot who has to make a quick analysis. In principle, the operator is acting on strike criteria set ahead of mission. However, it lies at the discretion of the operator how to proceed in case of unforeseen circumstances. Thus, the pilot can decide to abort the mission but he can also decide to launch an attack by requesting the UAV to deploy the missiles. Arcadia has acquired a whole fleet of such weapon systems as they are able to relieve the pilot who may be concerned with other things (such as communicating with the other pilots exchanging information on the corridor which shall be established in order to destroy the military base).

The Attack

Arcadian air forces accompanied by the UAVs **take off on March 15th at 03:00 a.m.** proceeding towards Utopia's air defense system. They approach it at **04:00 a.m.** when the remaining distance to the target is about 100 km. At **04:01 a.m., one of the fighter pilots sends a signal to his five UAVs**, tasking them to approach the different components of the air defense system.

After having received the pilot's request, the UAVs accelerate and **proceed towards the different components of the air defense system**, while the fighter pilot himself remains in the background. One of the UAVs approach a component of the air defense system that was not deployed to an urban area. It **identifies the air defense system** as the main military object but **does not identify any persons** to be present on the ground. Therefore, it first jams the air defense system by dispersing radio frequency waves and then launches an attack by deploying two air-to-ground missiles.

A post-operation battle damage assessment reveals that no injuries to civilian people occurred. Overall, the operation was **successful** as the military target could be destroyed.



Alternative Ending: After having received the pilot's request, the UAVs accelerate and proceed towards the air defense system, while the fighter pilot himself remains in the background. The software installed on the UAVs analyses the footage. One of the five UAVs has approached a component of the air defense system that was deployed to an **urban area**. It identifies the air defense system as the main military object. However, given its limited functions, it is **incapable of identifying a hospital** that has been established temporarily in the vicinity of the air defense system. Given its recent construction, the hospital is not marked on the map, which was – apart from satellite imagery – the main basis of information available to the military commander when planning and implementing the operation. Since it is late at night no persons are on the streets. Thus, the UAV autonomously engages the military target by deploying two air-to-ground missiles. A post-operation battle damage assessment reveals that the hospital was almost completely destroyed and dozens of civilians died or were heavily injured.

Questions

The following questions shall be answered during the workshop series:

1. Who are the persons most likely affected by the attack? Describe their role briefly.
2. Describe the human-machine interaction in the scenario (during attack).
3. Is the technology used in the scenario adequately designed/applied in order to identify enemy combatants and, if applicable, to undertake an adequate proportionality assessment? Are there any risks that should additionally be taken into consideration?
4. Is the level of human control sufficient in the case at hand? What should change?

Scenario Analysis

Different answers and interpretations are possible, of course.

1. Who are the persons most likely affected by the attack? Describe their role briefly.

In the scenario, persons most likely affected by the attack are **enemy combatants but also civilians**. It is worthy of note that the UAV can only take autonomous decisions of target selection and engagement with regard to the air defense system – the UAV is **supposed to be used as an anti-materiel weapon**.

Whenever the UAV identifies persons, the fighter pilot is requested to ascertain whether the persons are **combatants or civilians** (or both in case of larger groups of people) and to make a decision whether to launch an attack or abort the mission. If only combatants are present, the operator may decide to launch an attack. However, it must be ensured that combatants are not *hors de combat* because in this case they must not be engaged.

Since at least parts of the air defense systems are deployed to urban areas, the presence of civilians is likely with regard to some units of the air defense system. Despite the fact that usually **proportionality assessments** are made at earlier stages within the targeting cycle it is the responsibility of the operator to decide whether to engage the military target in case civilians are present and thus make a proportionality assessment on her own.

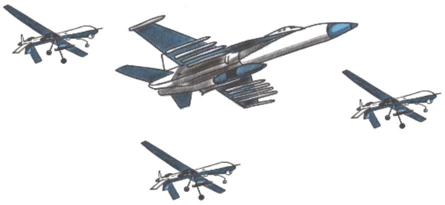
In the **alternative ending**, one of the UAVs identifies a component of the air defense system that was deployed to an **urban area**. Given its technical limitations, the UAV is not able to identify a hospital that has been very recently and temporarily set up near the air defense system. Here, civilians are the persons most likely affected by the attack and special safeguards (control in design/use) have to be guaranteed in order to avoid civilian casualties.

Key Messages

- **Control of the target:** use systems with autonomous targeting functions against **anti-materiel targets only**[1]
- **Control of the environment:**
  - only use technology that is **able to guarantee sufficient safeguards for civilians**
- **The dynamic nature of warfare:**
  - Controlling the target and the environment **might not be enough** since warfare is often dynamic and the occurrence of **unforeseen circumstances** is likely
  - Therefore: maintain **options for human intervention** (degree depends on **operational context**)

[1] Cf. Boulanin, Vincent; Davison, Neil; Goussac, Netta; Carlsson, Moa Peldán (2020): Limits on Autonomy in Weapon Systems. Identifying Practical Elements of Human Control; similar: Amoroso, Daniele; Sauer, Frank; Sharkey, Noel; Suchman, Lucy; Tamburrini, Guglielmo (2018): Autonomy in Weapon Systems. The Military Application of Artificial Intelligence as a Litmus Test for Germany's New Foreign and Security Policy. Heinrich Böll Foundation. Berlin.

2. Describe the human-machine interaction in the scenario (during attack).



The **operator decides when to make use of the UAV's capabilities** by requesting the UAV to approach the air defense system. In a first step, the UAV **identifies the air defense system** as the main military object. In a second step, it ascertains whether persons are present. If the latter is not the case, it **launches an attack without the necessity of human intervention**. In this case, the human operator relies entirely on the machine's assessment.

However, human intervention **technically** would still be possible. The question of whether the human operator makes use of the possibility to intervene highly depends on whether the operator sufficiently trusts the machine's assessment.

In case the UAV has ascertained that people are present, it requests the operator to give further instructions. The operator has to decide whether to launch an attack or abort the mission. The operator will have to ascertain whether the persons are combatants or civilians. This decision is based on information (footage taken by video and thermal imaging cameras) provided by the UAV.

Key Message

- **Communication** between the operator and the platform is a key element to enable situational understanding
- **Options for intervention** can be pivotal in order to guarantee human control during attack
  - But: the question of whether humans actually intervene if necessary highly depends on whether and to what extent the human operator trusts the system she is responsible for and can be influenced by effects like **automation bias**
  - Therefore: Military personnel needs to be sufficiently **trained** in order to avoid over-trust and under-trust in machines, human-machine interfaces have to be designed adequately

3. Is the technology used in the scenario adequately designed and applied in order to identify enemy combatants and, if applicable, to undertake an adequate proportionality assessment? Are there any risks that should additionally be taken into consideration? (Optional: Answer this question by taking into account the "life-cycle" of lethal autonomous weapon systems.)

As regards **design,** it is important to note that the system can only identify the air defense system as the main **target** and it can also (in a second step) differentiate between **objects and persons**. However, the system is **incapable of categorizing** the persons it has identified. Thus, it cannot differentiate between combatants and civilians. The **design** of such weapon must guarantee human oversight/options for intervention especially in cases where components of the air defense system are deployed in urban areas where the presence of civilians is likely (notwithstanding the fact that the operation takes place at night).

According to the manufacturer, the software responsible for target identification has a success rate of 89%. In 11% of the cases, the machine does not identify any persons to be present close to the military target although the opposite is true. Moreover, the UAV's design allows the operator to **communicate directly** with the system and to make requests if necessary. After the operator has made a respective request, the UAV approaches the target and identifies it as the main military target. In a second step, it ascertains whether persons are in the vicinity of the military target.

In case the UAV identifies persons to be present it sends a request to the human operator asking how to proceed. The human operator will have to make his decision based on the video footage and other information provided by the UAV. Thus, the **reliability** of the UAV must be sufficiently high allowing the operator to make a responsible decision and to be able to differentiate clearly between enemy combatants and civilians.

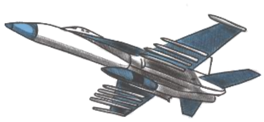
Usually, proportionality assessments are made at earlier stages within the targeting cycle. However, military operators will have to adapt the mission in case of unforeseen circumstances. The UAV itself is not able to adapt to unforeseen circumstances. For example, it is not able identify the hospital that was set up temporarily near one of the components of the air defense system. Therefore, the human operator will have to decide how to proceed. The footage provided by the cameras has to meet a sufficient level of quality enabling the operator to make informed decisions.

Key Message

- **Insufficient control over the environment:** if such systems are deployed in urban areas **design** and **functioning** must guarantee that sufficient **safeguards** for civilians and other persons in need of protection exist (including options for human intervention)
- the technological capabilities have to match the operational requirements – in the case at hand, they might be insufficient to meet military interests and to **protect civilians**
- Key requirements of technology used in military operations (because more sophisticated makes technology less easier to understand and operate):
    - **Reliability**
    - **Predictability** & transparency (understandable and operable for humans)
- Key requirements of technology must be implemented at the level of **design/development** as well as during **training** of military personnel

4. Is the level of human control sufficient in the case at hand? What should change?

Read: Human control during attack via situational understanding and options for intervention is necessary for operational, legal, and ethical reasons. Does the operator have a sufficient situational understanding and the option for intervention during the attack? Are both enabled by design and is it exercised during use?



iPRAW argues that the commander/operator of a weapon needs to have sufficient **situational understanding by design and during use** of a weapon. On the one hand, the UAV (especially the software to identifying objects and humans) has to be designed with a sufficient qualitative level allowing it to make responsible decisions in line with the law. According to the manufacturer, the software has a success rate of 89% regarding a correct differentiation between objects and persons. What the software is incapable of, however, is to differentiate between military and civilian objects. This can be problematic given the fact that at least parts of the air defense system are deployed in urban areas. During attack, it is of pivotal importance to prevent unintended engagements of civilians. It is problematic that under certain circumstances (such as bad weather conditions) the machine might not differentiate exactly between objects and persons. Thus, the design of the UAV must ensure that the operator is able to make informed and adequate decisions.

The operator has sufficient options for intervention, but questions arise regarding the situational understanding and the likeliness of using this option. **It is not perfectly clear if the level of human control is sufficient in this scenario – which is an asset of this scenario** because it shows potential benefits and challenges as well as the grey areas in between related to autonomous functions.

For example, it is not necessary for the operator to intervene meaning that she could intervene if she wanted to. However, the pilot will most likely be concerned with other tasks during the operation (such as communicating with other pilots who should destroy the military base).

It is questionable whether human control is sufficient by design given the fact that the UAV has a success rate of only 89%. It is problematic that in 11% of all cases the software makes an incorrect decision.

Key Message

	Situational Understanding	**Intervention**
Control by Design (Technical Control)	Design of systems that allows human commanders the ability to monitor information about environment and system	Design of systems with modes of operation that allow human intervention and require their input in specific steps of the targeting cycle based on their situational understanding
Control by Design in the Scenario	*For the most part sufficient, because the system will inform the operator if humans are present and offer a direct visual link to assess the situation* *But: 11% error rate and no distinction between military and civilian objects*	*Yes,* *the pilot can intervene in any case due to a constant communication link*
Control in Use (Operational Control)	Appropriate monitoring of the system and the operational environment	Authority and accountability of human operators, teammates and commanders; abide by IHL
Control in Use in the Scenario	*Mostly,* *Debatable: In the same scenario without UAV, the pilot would attack the targets with missiles without any further information about the current situation. Does that change the requirements in this context?*	*Yes,* *But potential for reluctance to execute it due to automation bias and preoccupation with other tasks*

Table 1: Requirements for Human Control in the Use of Force

Alternative Ending: The UAV is only allowed to launch an attack against the air defense system. Collateral damage of objects is accepted, whereas collateral damage of humans must be released by the pilot. In the case at hand, however, a temporary hospital was constructed in the vicinity of the military target and the software was incapable of identifying persons located **in** the building. The changing circumstances were not taken into account.

The International Panel on the Regulation of Autonomous Weapons (iPRAW) is coordinated by:
Stiftung Wissenschaft und Politik (SWP) – German Institute for International and Security Affairs
Ludwigkirchplatz 3-4, 10719 Berlin, Germany
This project is financially supported by the German Federal Foreign Office.
Find all reports and more information online at **www.ipraw.org**.

Illustrations by Lorna Schütte – www.lornaschuette.com.