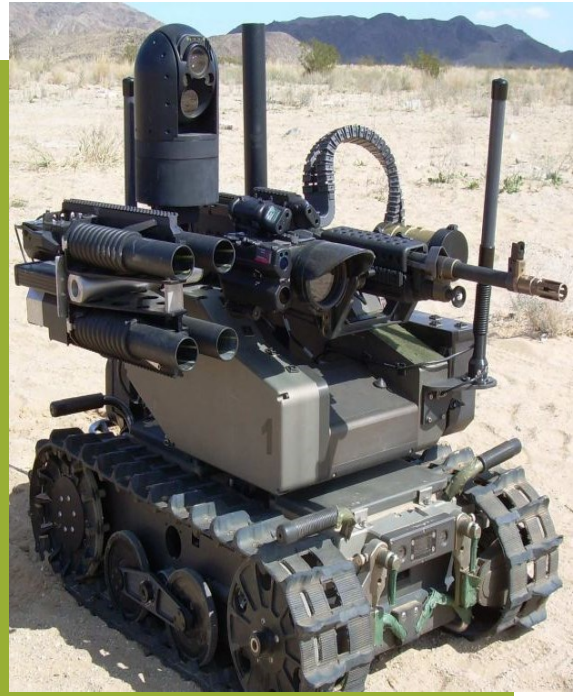


Human supervisory control of autonomous weapons

Francesco Farinola
Michele Vece

Table of contents

"Towards a principle for the human supervisory control of robot weapons"

Sharkey, 2014

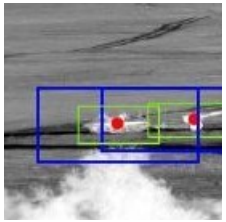
01

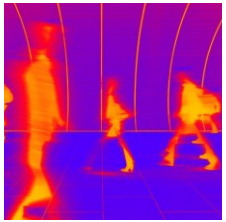
Automatic vs aided target recognition

Target recognition methods: why...

shape detection

makes it possible to **recognize objects** in an uncluttered environment





thermal imaging

detects **heat radiating** from an object and shows its **movement**

radiation detection

used by loitering munition to detect **radar signal** and determine if they are friendly

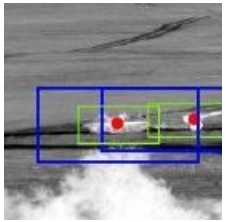
acoustic direction finding

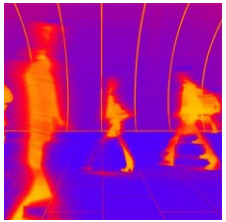
calculates the **location** of the sound by using differences between the times that sound reaches two or more separated microphones

… and why not!

shape detection

medium to high **cluttered environments** introduce an unacceptably high **false alarm rate**





thermal imaging

could not be used to distinguish between a **combatant** and **civilians**.

radiation detection

the radar should be part of an anti-aircraft installation in order to **determine the legitimacy** of a target.

acoustic direction finding

other acoustic effects may be **detected** and **responded to**

*"Autonomous and semi-autonomous weapon systems shall be designed to allow commanders and operators to exercise appropriate levels of **human judgment** over the use of force"*

— US Department of Defence

*"Fully autonomous systems rely on a certain level of artificial intelligence for making high-level decisions from a very complex environmental input, the result of which might not be fully predictable at a very detailed level. However, let us be absolutely clear that the operation of weapons systems will always be **under human control**"*

— J. Astor

Conclusions

Even an improved ability to recognize targets does **not** enable machines to assess whether a target is **legitimate** and whether the attack as a whole is permissible.

Thus, a more appropriate use of these methods would be to **assist** human **supervisory control** to achieve more precise and accurate targeting by a human exerting the power of **deliberative reasoning** and **judgement**.

02

The delicate human and computer balancing act

The human and computer balancing act

Because humans sometimes fail at some tasks, it does **not** mean that machines can do them any better.
It can simply mean that humans are being asked to perform in a mode of operation that is not well suited to **human psychology**.
This needs to be part of the equation of ensuring **efficient** and **meaningful human supervisory** control of weapons.

If we get the balancing act right we could have **more precision** and **accurate targeting** with **less collateral damage** and **better predictable compliance** with International Humanitarian Law.
But getting it wrong could result in considerable **humanitarian problems**.

Deliberative vs automatic process

Deliberative reasoning

- «Your conscious self – the thinking you». Kahneman D.
- Goes along with the automatic processes unless there is something surprising or irregular and/or we are operating in **novel circumstances** or performing tasks that require **vigilance** and/or **deliberation**.
- Requires **attention** and free **memory space**
- Stress or distractions could incapacitate it

Automatic reasoning

- Does **not** require active control or attention
- Can be trained through **repetition** and **practice** on **routine tasks**.
- Used anytime for routine decisions that have to be made **rapidly** for **predictable events**
- Works well in environment that contains useful **cues** that, via practice, have been (over) rehearsed.

Does a given domain afford enough **regularity** to be **learnable** as an **automatic process**?

Why automatic reasoning is problematic (1)

01

It neglets ambiguity and suppresses doubt

- Automatic processes are all about jumping to conclusions.
- They are guided by **experience**.
- An unambiguous answer pops up immediately and does not allow doubt.
- It does not search for **alternative interpretations** and does not examine **uncertainty**.

02

It infers and invents causes and intentions

- It is adept at finding a **coherent causal story** to link together fragments of available information.
- **assimilation bias:** a person who is presented with new information that contradicts a preexisting mental model, assimilates the new information to fit into that mental model.

Why automatic reasoning is problematic (2)

03

It is biased to believe and confirm

- It favors the **uncritical acceptance** of suggestions and maintains a **strong bias**
- **Automation bias:** a computer-generated solution is accepted as correct, human decision maker disregards or does not search for contradictory information
  - **Error of omission**: humans fail to notice problems because the automation does not alert them
  - **Error of commission**: humans erroneously follow automated directives or recommendations
- **Confirmation bias:** when seek out information to confirm a prior belief and discount information that does not support this belief

04

It focuses on existing evidence and ignores absence evidence

- It builds a coherent explanatory history without considering any **evidence** or **contextual information** that might be **missing**
- **WYSIATI:** «What You See Is All There Is»
- It makes us confident to accept information as true whether it is or not

Conclusions

The unpredictable and unanticipated circumstances in a **dynamically changing environment** play to the **weakness** of automatic reasoning.

It is vitally important that **deliberative reasoning** is enabled in the design of **supervisory control** for weapons systems.
Although this is also subject to error and flaws, it does as good a job as can be done with **uncertainty** and **doubt**.

If a supervisory weapons operator is distracted by another task or if they are stressed, their **attentional capacity** may be low. They would not be able to use their **deliberative reasoning** and could simply catch the downsides of **automatic reasoning** if there were problems or irregularities.

03

Deliberative reasoning meets supervisory control

To say that a “**human is in-the-loop**” does not clarify the degree of **human involvement**

Levels of human supervisory control (1)

1. human deliberates about a target before initiating any attack
2. program provides a list of targets and human chooses which to attack
3. program selects target and human must approve before attack
4. program selects target and human has restricted time to veto
5. Program selects target and initiates attack without human involvement

Levels of human supervisory control (2)

01

Acceptable*

* if requirements of deliberative human control met:

- ❑ full contextual and situational **awareness** of the target area
- ❑ be able to perceive and react to any **change** or **unanticipated** situations
- ❑ **active cognitive participation**
- ❑ **sufficient time for deliberation** on the nature of the target, its significance in terms of the necessity and appropriateness of attack
- ❑ means for the **rapid suspension** or **abortion** of the attack.

02

Acceptable*

* if requirements of deliberative human control met and no **ordered list** is provided, otherwise there would be a tendency to accept the top ranked target unless sufficient time and attentional space is given for deliberative reasoning

Levels of human supervisory control (3)

03

Unacceptable

Creates an **automation bias** in which human operators come to accept computer generated solutions as correct and disregard or don't search for contradictory information (Cummings, 2004)

04

Unacceptable

- It does **not** promote **target identification**
- Short time to veto would reinforce **automation bias** and leave no room for doubt or deliberation.
- The **time pressure** will result in operators falling foul of the four **downsides** of automatic reasoning

05

Unacceptable

- No **human involvement** in the target selection and attack.
- No **compliance** with international law

04

Human supervised autonomy?

SARMO weapons

- **SARMO** (Sense and React to Military Objects)
- Weapons systems in use that operate **automatically** once activated: they intercept high-speed **inanimate** objects automatically.
- They complete their **detection**, **evaluation** and **response** process within a matter of seconds
- **Difficult** for human operators to exercise **meaningful supervisory control** once they have been activated other than deciding when to **switch them off**.
- Precursors to **fully autonomous** weapons (Human Rights Watch, 2012)

SARMO weapons: features

- small set of **defined actions** repeatedly and independently of **external influence** or control
- highly **structured** and **predictable environments**
- switched on after detection of a **specific threat**
- **unable** to dynamically initiate a **new targeting goal** or **change mode** of operation
- constant **vigilant human evaluation** and **monitoring** for **rapid shutdown**
- **predictable** output and behaviour
- only used **defensively** against **direct attacks** by military objects

SARMO weapons: be cautious!

- From the perspective of the human supervisory control framework, the **human decision** of when to use the weapon is the **key** to the **legality** of SARMO weapons systems.
- It is essential that **precautionary measures** have been taken

*"The potential damage caused by not using C-RAM in its automatic mode **justifies** the level of any anticipated **collateral damage**"*

— UK Ministry of Defence

- This omits **precaution**, **proportionality** and **necessity**
- **Unacceptable** under International Humanitarian Law
- **Incautious** use of unsupervised weapons could cause **disproportionate harm** to civilian populations and objects

SARMO weapons: erosion

*"The role of the human in the loop has, before now, been a legal requirement which we now see being **eroded**"*

— UK Ministry of Defence

- **Avoid** such **erosion**
- **Lock down** human supervisory control as a legal principle of human control

*«MANTIS' control system is also capable of tracking the location of the **assailants** along with the flight path and point of impact»*

— MANTIS manufacturer's specification

- **Do not assume** that the assailants are present at the location!
- It is up to the commander to assess whether or not there are **legitimate** targets at the location

Overall conclusions

- **Fully autonomous** weapons **could not be used** in a way that could be guaranteed to predictably comply with International Law.
- There is general agreement on the **inadequacy** of Automatic Target Recognition
- Both humans and computer systems have their strengths and weaknesses: **exploit** the **strengths** of **both!**
- Do **not** ask humans to perform in a mode not well suited to **human psychology**
- Do **not** use computerized weapons that are not **meaningfully controlled** by human operators.
- Develop a **principle** of human control founded on **human reasoning processes** to provide clear **guidelines** for state weapons reviews.

What type of **human control** will be employed?
How **meaningful** it will be?