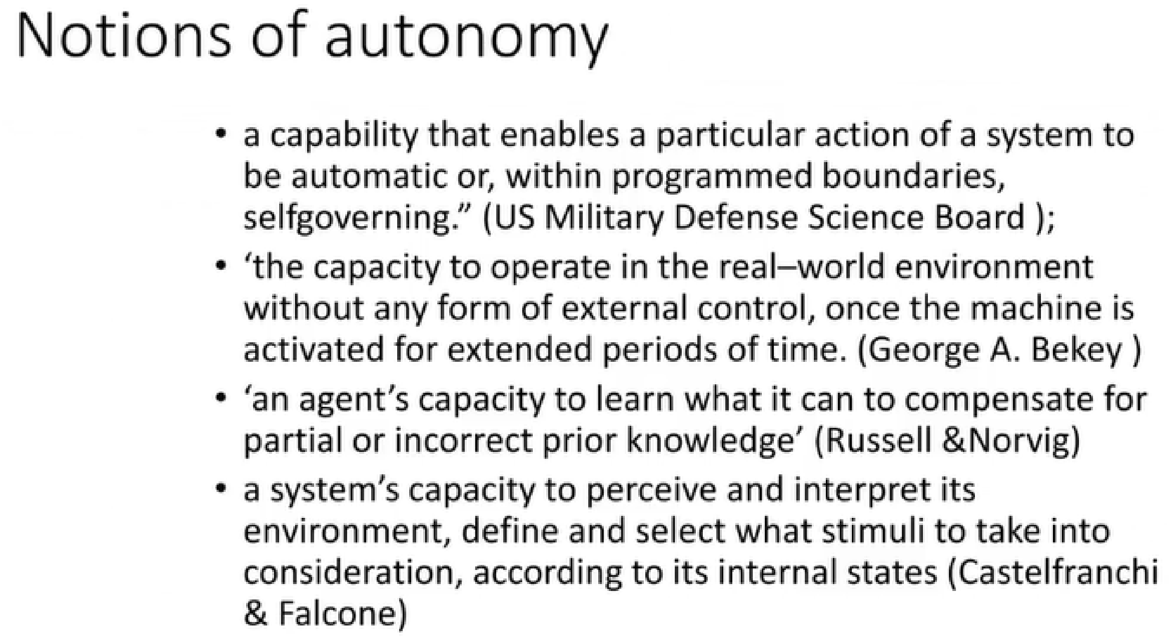
Lesson_12_Intelligent_Weapons

Introduction and Features of autonomous systems

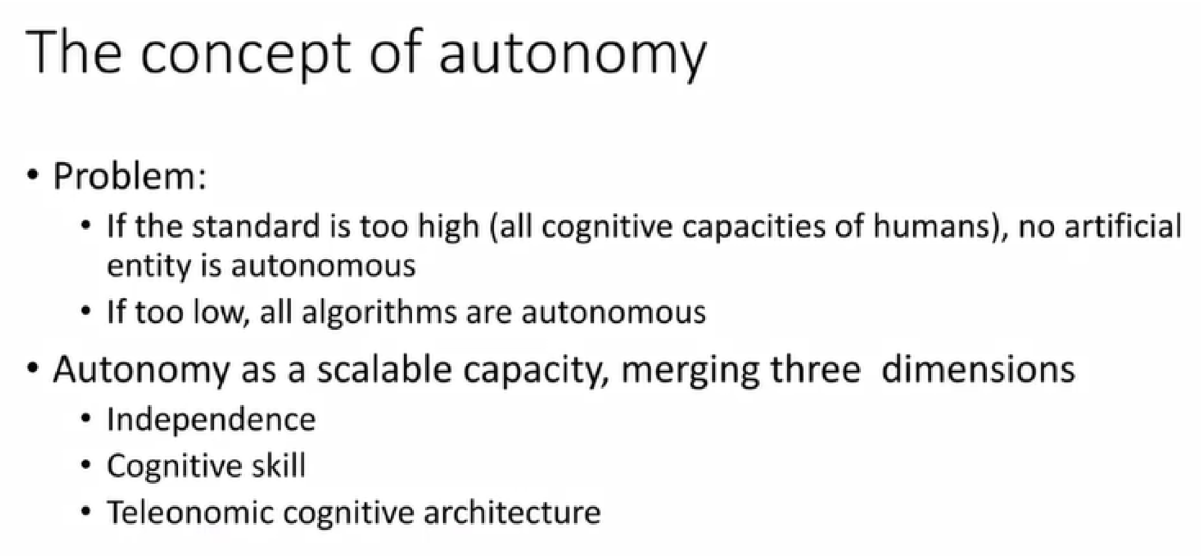


Notions of autonomy

- a capability that enables a particular action of a system to be automatic or, within programmed boundaries, selfgoverning.” (US Military Defense Science Board);
- ‘the capacity to operate in the real–world environment without any form of external control, once the machine is activated for extended periods of time. (George A. Bekey)
- ‘an agent’s capacity to learn what it can to compensate for partial or incorrect prior knowledge’ (Russell &Norvig)
- a system’s capacity to perceive and interpret its environment, define and select what stimuli to take into consideration, according to its internal states (Castelfranchi & Falcone)

The notion of autonomy is key when addressing the issue of autonomous weapons. When an automatic system is to be considered **autonomous**? A door that opens on its own using a proximity sensor is autonomous? Or should it be capable to pursue goals and find a way to achieve them?

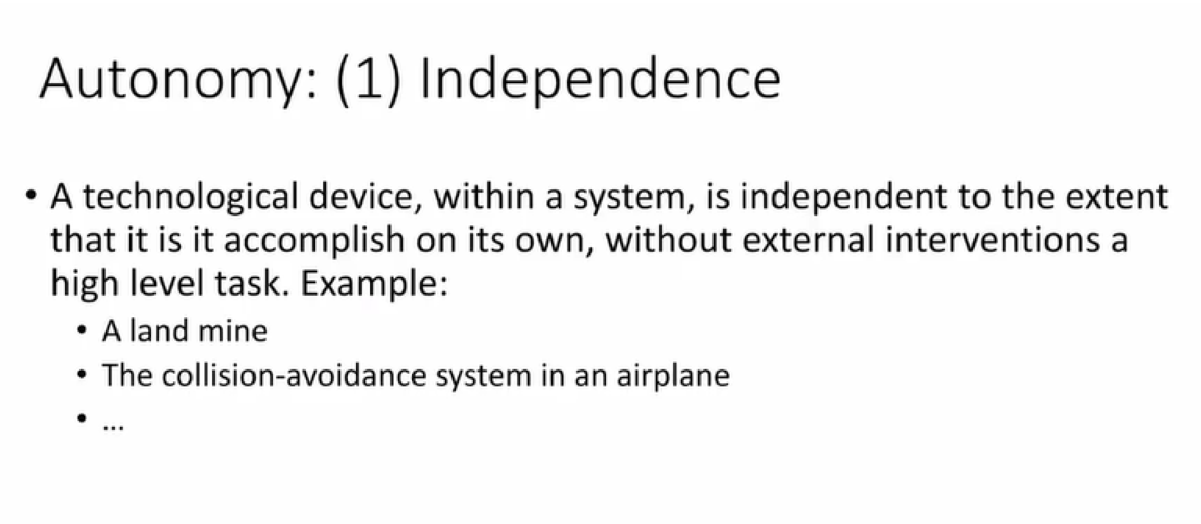
Autonomy can be conceived in different ways.



The concept of autonomy

- Problem:
  - If the standard is too high (all cognitive capacities of humans), no artificial entity is autonomous
  - If too low, all algorithms are autonomous
- Autonomy as a scalable capacity, merging three dimensions
  - Independence
  - Cognitive skill
  - Teleonomic cognitive architecture

A *Teleonomic cognitive architecture* allows the entity to purse its goals.



Autonomy: (1) Independence

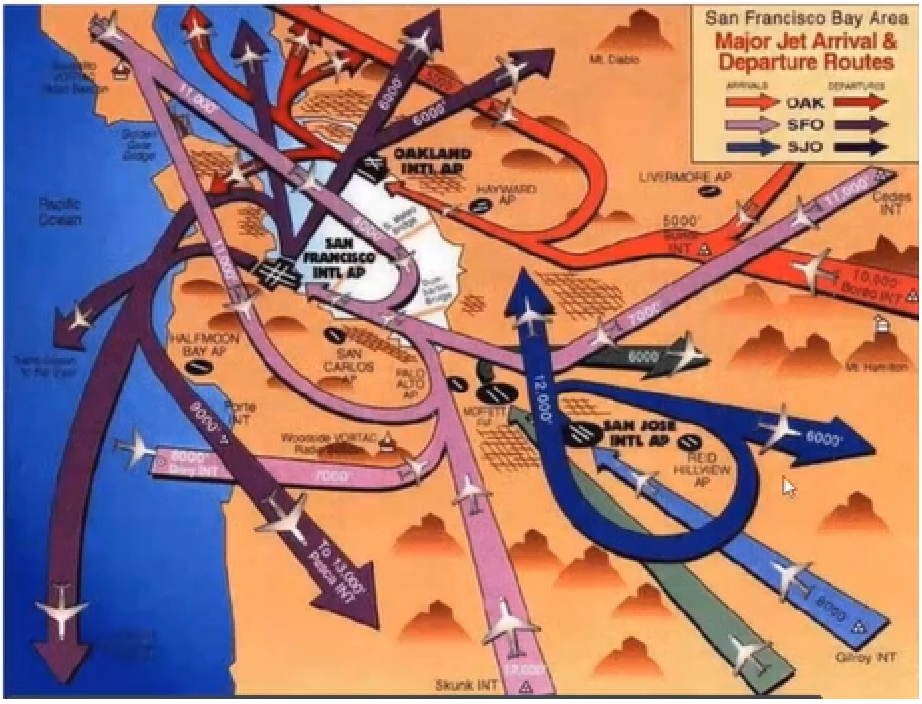
- A technological device, within a system, is independent to the extent that it is it accomplish on its own, without external interventions a high level task. Example:
  - A land mine
  - The collision-avoidance system in an airplane
  - …

Indipendence is related to the idea of **automation**, a device is considered **independent** when it can accomplish its own task without external interventions.

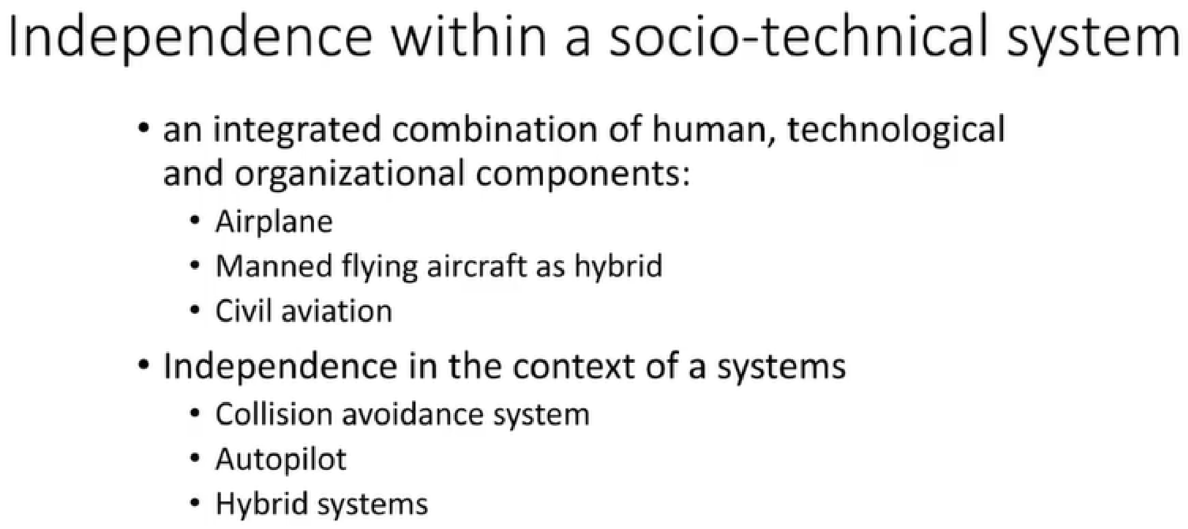
In the military domain there are some devices (e.g., land mine, their use is now banned) that respect this paradigm; also collision-avoidance systems work *indepentently*, steering airplanes that might collide without human intervention.

There are various degrees of independence: just think about airplanes' evolution, where initially all the task had to be carried out by humans and now we have drones that work on their own, from the take off to possibly shooting missile that reach their target autonomously.

Aviation system



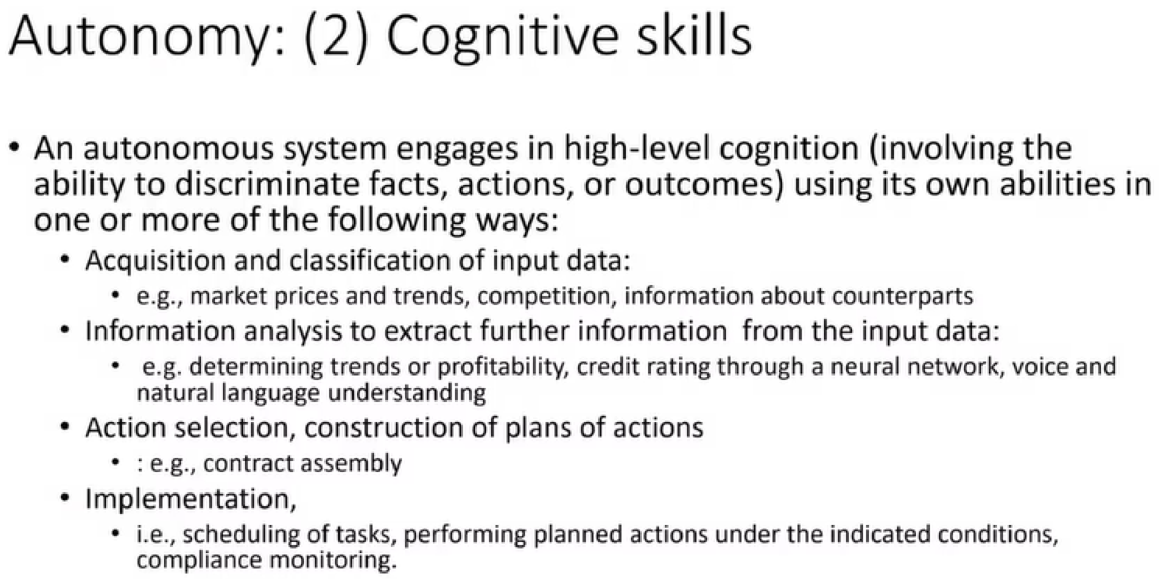
Complex Aviation system where automation is fundamental for its management



Independence within a socio-technical system

- an integrated combination of human, technological and organizational components:
  - Airplane
  - Manned flying aircraft as hybrid
  - Civil aviation
- Independence in the context of a systems
  - Collision avoidance system
  - Autopilot
  - Hybrid systems

Example of how autonomous independent components can be integrated in a complex system



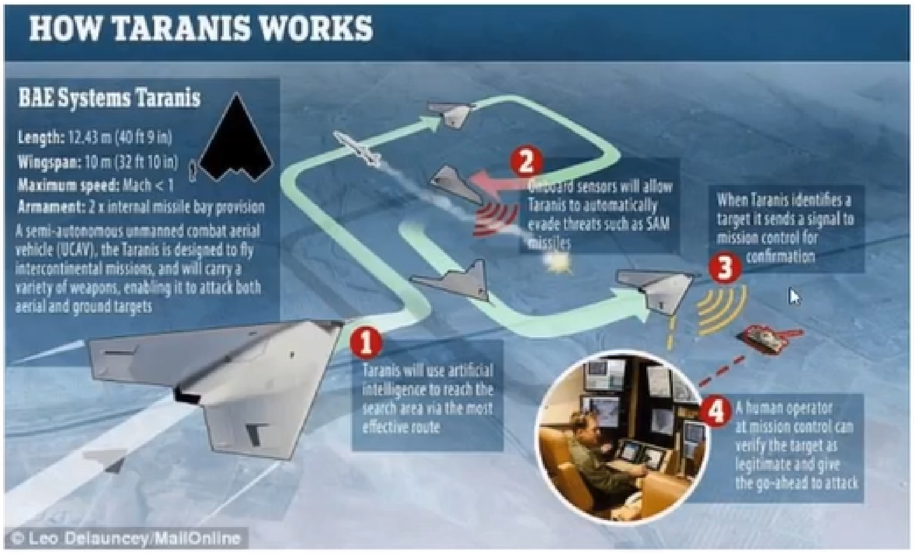
Autonomy: (2) Cognitive skills

- An autonomous system engages in high-level cognition (involving the ability to discriminate facts, actions, or outcomes) using its own abilities in one or more of the following ways:
  - Acquisition and classification of input data:
    - e.g., market prices and trends, competition, information about counterparts
  - Information analysis to extract further information from the input data:
    - e.g. determining trends or profitability, credit rating through a neural network, voice and natural language understanding
  - Action selection, construction of plans of actions
    - : e.g., contract assembly
  - Implementation,
    - i.e., scheduling of tasks, performing planned actions under the indicated conditions, compliance monitoring.

To speak of autonomy, we need to talk about **cognitive skills.** A land mine works **independently**, someone steps on it and boom goes the dynamite, but autonomy is reached only when the system is able to engage to some **high-level cognition task**, rather to just reacting to an environmental change (e.g, pressure on the trigger), and is able to discriminate actions, outcomes, input data etc (*slide)*. A land mine is an **independent** device, but it's not considered **autonomous**.

For example, a drone should be able to acquire information on the environment (wind strength, its own position and so on), build predictions on the acquired data (weather forecast), construct a plan of actions (the route) and the implement that plan.

An independent and autonomous system?



HOW TARANIS WORKS

BAE Systems Taranis

Length: 12.43 m (40 ft 9 in)
Wingspan: 10 m (32 ft 10 in)
Maximum speed: Mach < 1
Armament: 2 x internal missile bay provision

A semi-autonomous unmanned combat aerial vehicle (UCAV), the Taranis is designed to fly intercontinental missions, and will carry a variety of weapons, enabling it to attack both aerial and ground targets

1 Taranis will use artificial intelligence to reach the search area via the most effective route

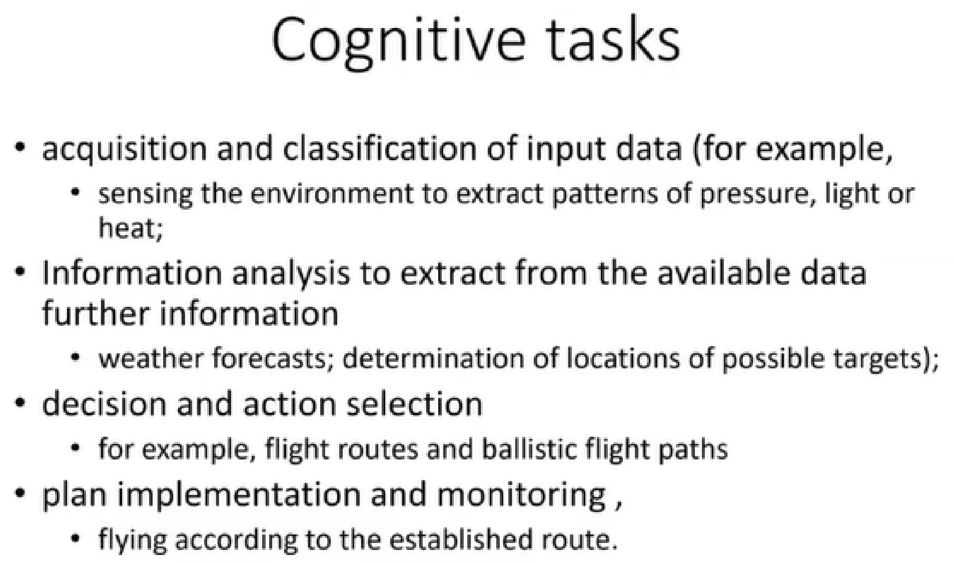
2 Onboard sensors will allow Taranis to automatically evade threats such as SAM missiles

3 When Taranis identifies a target it sends a signal to mission control for confirmation

4 A human operator at mission control can verify the target as legitimate and give the go-ahead to attack

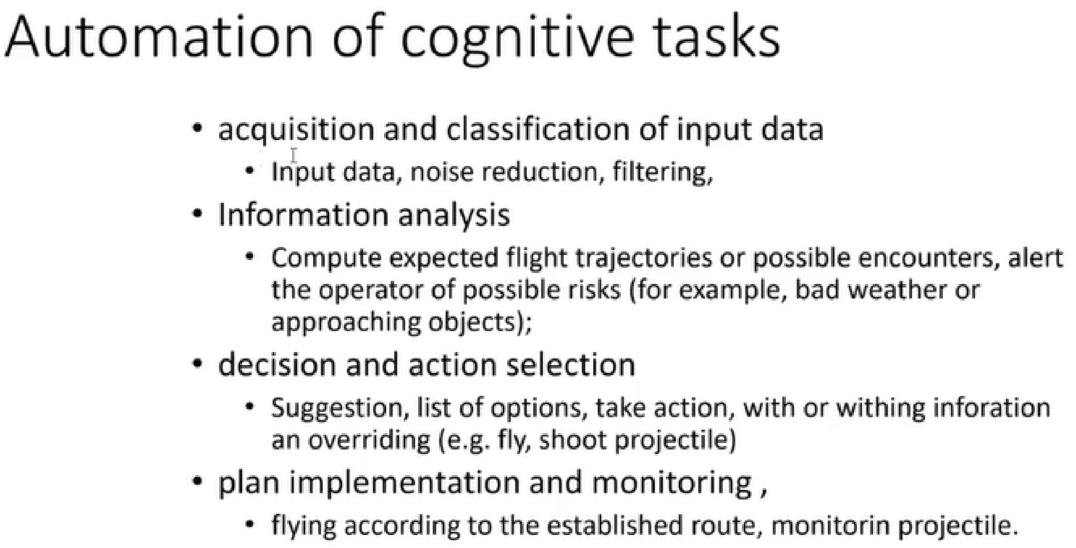
© Leo Delauncey/MailOnline

Even if there is a human in the loop, we can consider the system above as independent and autonomous as it respects all the requirements and the human is there only to verify the target and give authorization to fire.



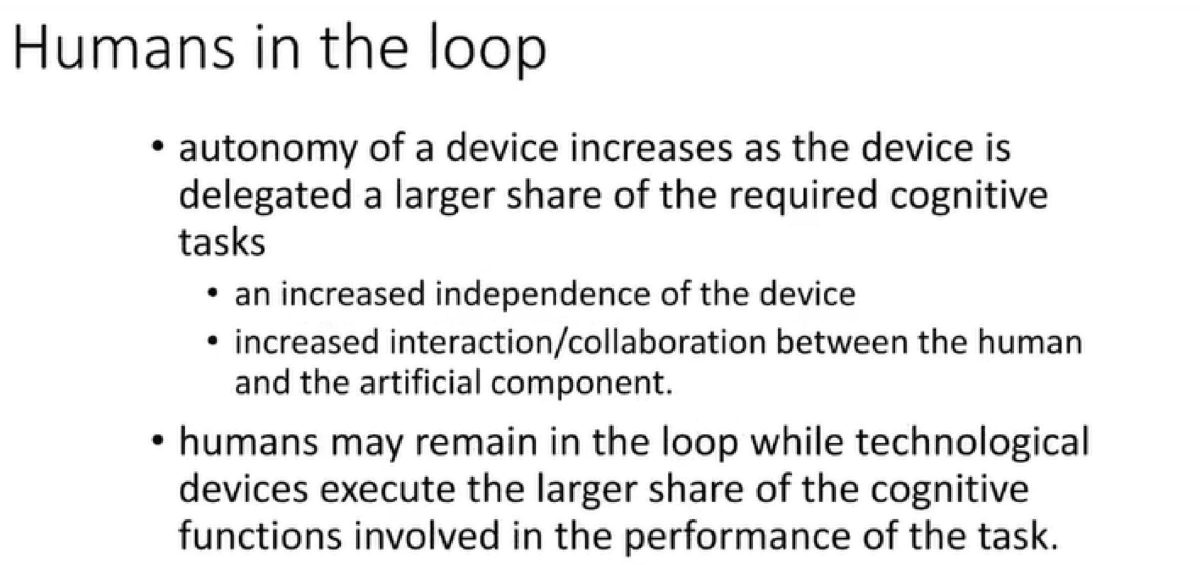
Cognitive tasks

- acquisition and classification of input data (for example,
  - sensing the environment to extract patterns of pressure, light or heat;
- Information analysis to extract from the available data further information
  - weather forecasts; determination of locations of possible targets);
- decision and action selection
  - for example, flight routes and ballistic flight paths
- plan implementation and monitoring ,
  - flying according to the established route.



Automation of cognitive tasks

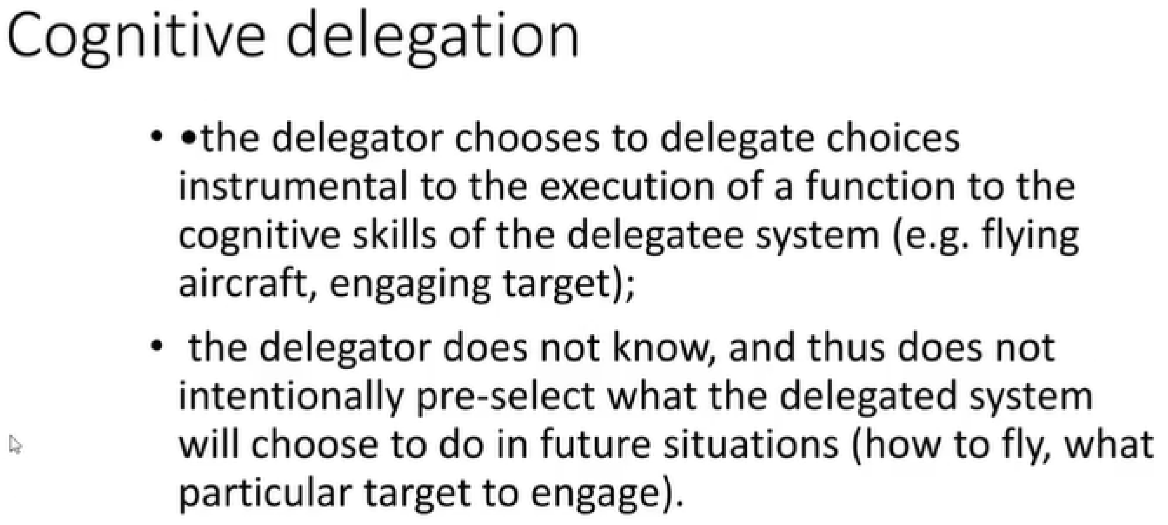
- acquisition and classification of input data
  - Input data, noise reduction, filtering,
- Information analysis
  - Compute expected flight trajectories or possible encounters, alert the operator of possible risks (for example, bad weather or approaching objects);
- decision and action selection
  - Suggestion, list of options, take action, with or withing inforation an overriding (e.g. fly, shoot projectile)
- plan implementation and monitoring ,
  - flying according to the established route, monitorin projectile.



Humans in the loop

- autonomy of a device increases as the device is delegated a larger share of the required cognitive tasks
  - an increased independence of the device
  - increased interaction/collaboration between the human and the artificial component.
- humans may remain in the loop while technological devices execute the larger share of the cognitive functions involved in the performance of the task.

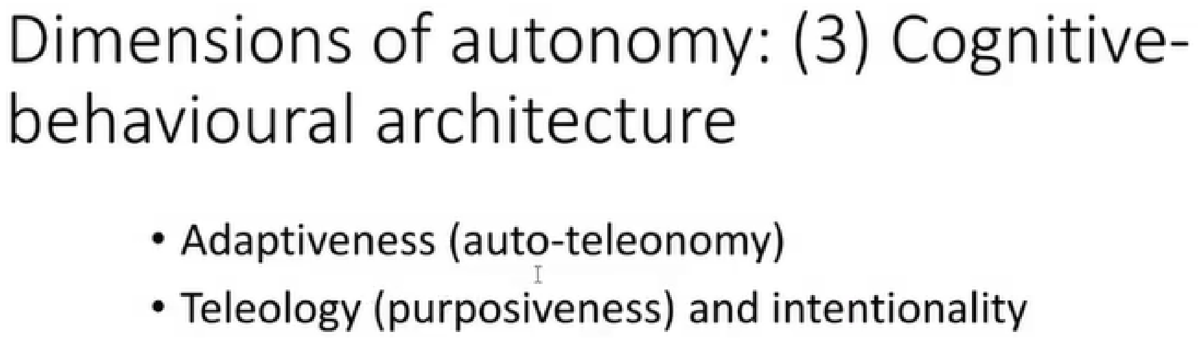
What kind of influence on the task is left to humans when they are on the loop?



Cognitive delegation

- •the delegator chooses to delegate choices instrumental to the execution of a function to the cognitive skills of the delegatee system (e.g. flying aircraft, engaging target);
- the delegator does not know, and thus does not intentionally pre-select what the delegated system will choose to do in future situations (how to fly, what particular target to engage).

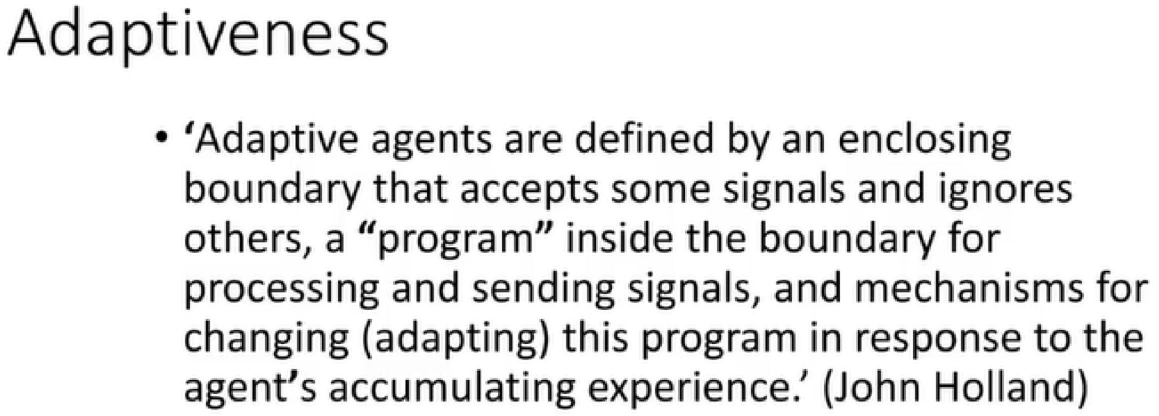
So, we're delegating to a tool not just material actions but also many cognitive functions. The human deploys such devices in order to achieve some kind of goal, but s/he does not decide the specific actions or cognitive operations done by the system.



Dimensions of autonomy: (3) Cognitive-behavioural architecture

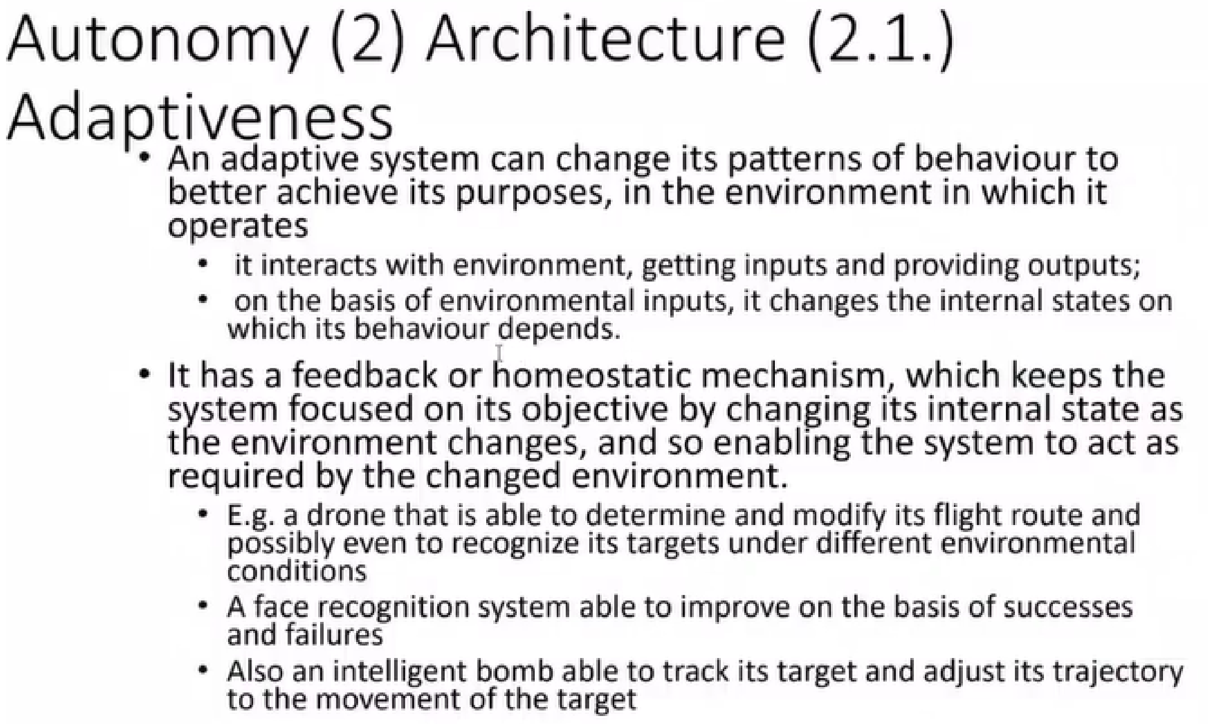
- Adaptiveness (auto-teleonomy)
- Teleology (purposiveness) and intentionality

- **Adaptiveness**: the system is able to adapt to the environment and can change the internal states accordingly in order to achieve its goal.



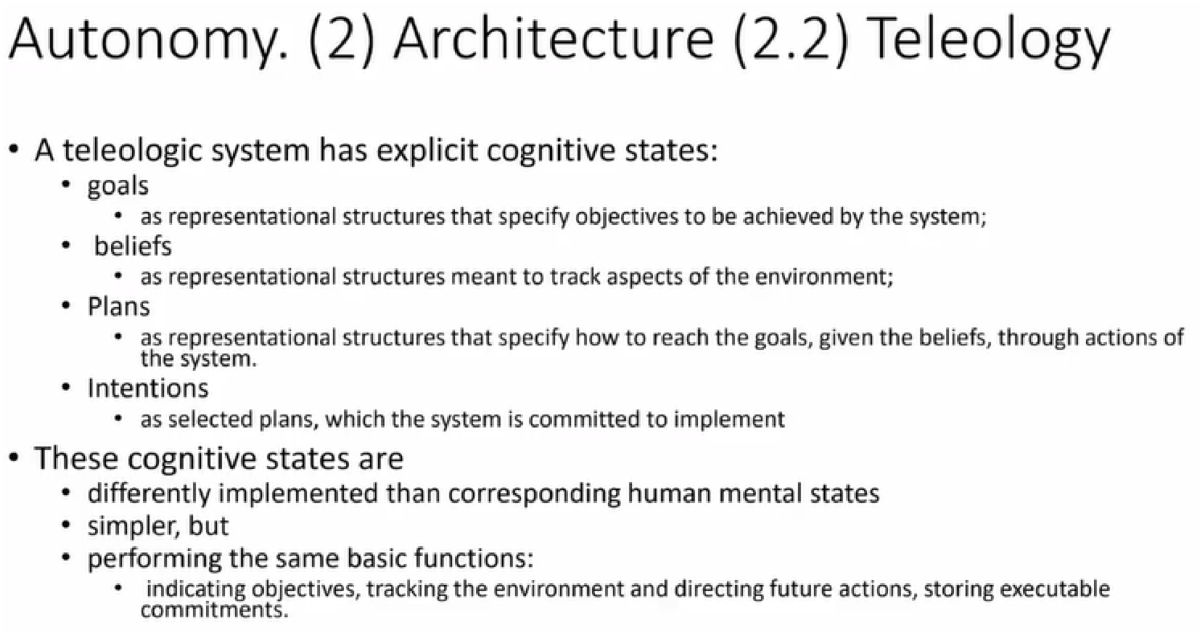
Adaptiveness

- 'Adaptive agents are defined by an enclosing boundary that accepts some signals and ignores others, a "program" inside the boundary for processing and sending signals, and mechanisms for changing (adapting) this program in response to the agent's accumulating experience.' (John Holland)



Autonomy (2) Architecture (2.1.) Adaptiveness

- An adaptive system can change its patterns of behaviour to better achieve its purposes, in the environment in which it operates
  - it interacts with environment, getting inputs and providing outputs;
  - on the basis of environmental inputs, it changes the internal states on which its behaviour depends.
- It has a feedback or homeostatic mechanism, which keeps the system focused on its objective by changing its internal state as the environment changes, and so enabling the system to act as required by the changed environment.
  - E.g. a drone that is able to determine and modify its flight route and possibly even to recognize its targets under different environmental conditions
  - A face recognition system able to improve on the basis of successes and failures
  - Also an intelligent bomb able to track its target and adjust its trajectory to the movement of the target

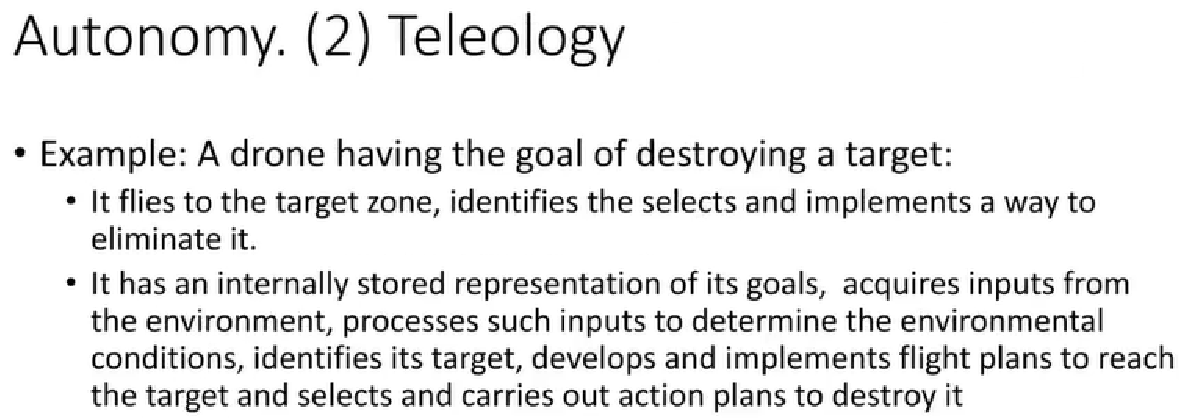


Autonomy. (2) Architecture (2.2) Teleology

- A teleologic system has explicit cognitive states:
  - goals
    - as representational structures that specify objectives to be achieved by the system;
  - beliefs
    - as representational structures meant to track aspects of the environment;
  - Plans
    - as representational structures that specify how to reach the goals, given the beliefs, through actions of the system.
  - Intentions
    - as selected plans, which the system is committed to implement
- These cognitive states are
  - differently implemented than corresponding human mental states
  - simpler, but
  - performing the same basic functions:
    - indicating objectives, tracking the environment and directing future actions, storing executable commitments.

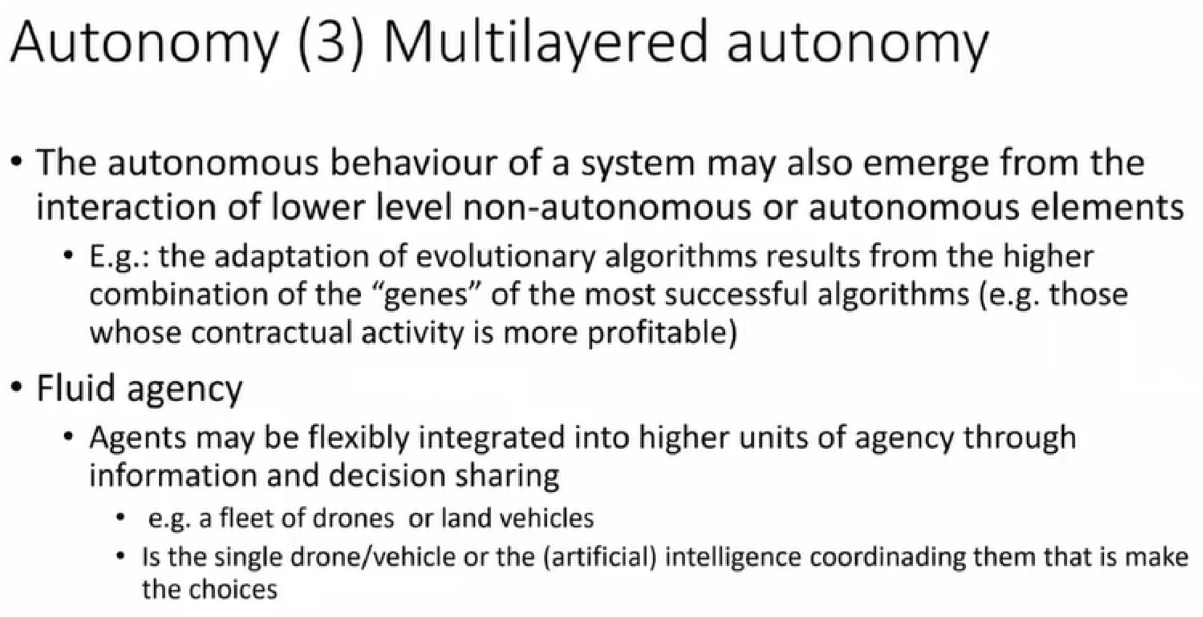
Teleology is in the domain of cognitive systems that are able to pursue goals and they have a cognitive states parallel to the human ones, even though they are not identical (obv).

We can develop theory of mind that can be applicable to entities that are not human minds but perform the same basic functions (Castelfranchi) .



Autonomy. (2) Teleology

- Example: A drone having the goal of destroying a target:
  - It flies to the target zone, identifies the selects and implements a way to eliminate it.
  - It has an internally stored representation of its goals, acquires inputs from the environment, processes such inputs to determine the environmental conditions, identifies its target, develops and implements flight plans to reach the target and selects and carries out action plans to destroy it



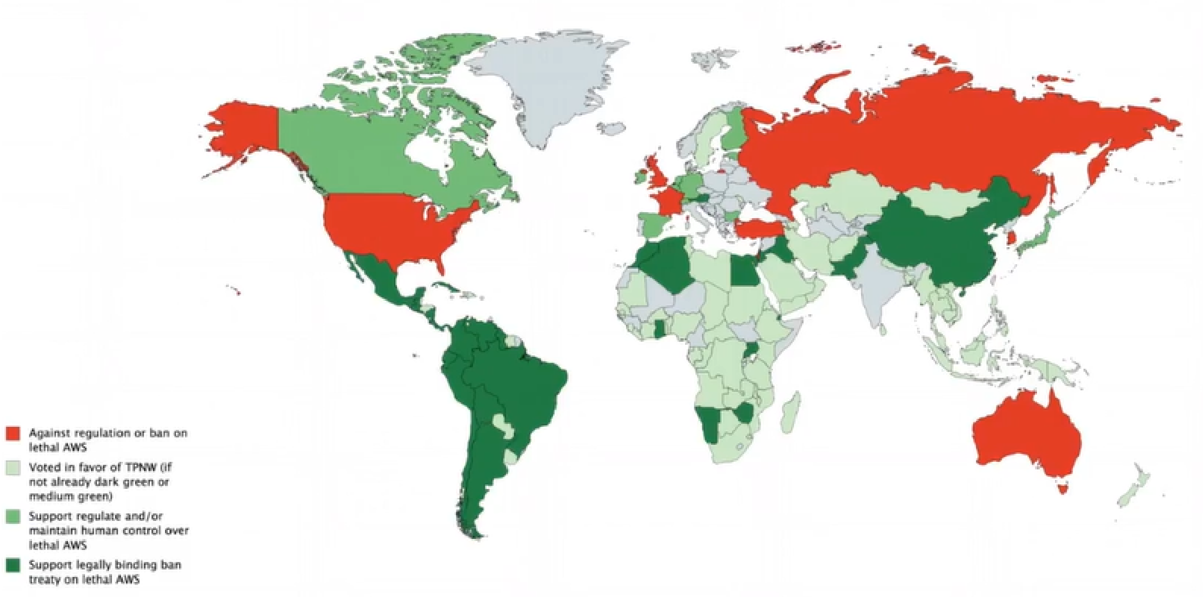
Autonomy (3) Multilayered autonomy

- The autonomous behaviour of a system may also emerge from the interaction of lower level non-autonomous or autonomous elements
  - E.g.: the adaptation of evolutionary algorithms results from the higher combination of the "genes" of the most successful algorithms (e.g. those whose contractual activity is more profitable)
- Fluid agency
  - Agents may be flexibly integrated into higher units of agency through information and decision sharing
    - e.g. a fleet of drones or land vehicles
    - Is the single drone/vehicle or the (artificial) intelligence coordinading them that is make the choices

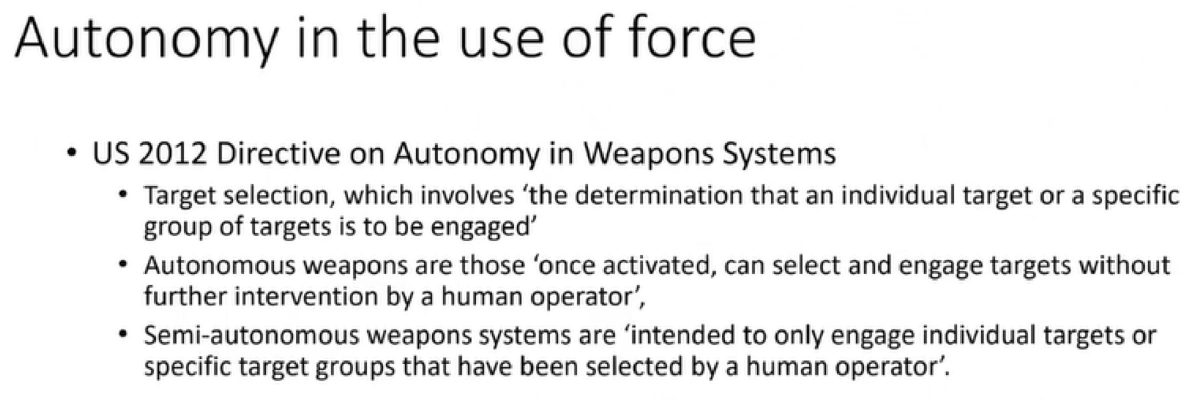
The idea that autonomy can arise from simpler and non-autonomous components was a key insight by the AI scholar Marvin Minsky.

Autonomy in weapon systems

There are differents attitudes wrt autonomous weapons across the world. Each country has its own interest in the issue based on their military and technological advancement.


🟥 Against regulation or ban on lethal AWS
🟩 Voted in favor of TPNW (if not already dark green or medium green)
🟩 Support regulate and/or maintain human control over lethal AWS
🟩 Support legally binding ban treaty on lethal AWS

Some countries (e.g, Germany) and Non-Governamental Organizations are pushing to ban, at least temporarily, the research and development of AW, but so far the debate creates huge division throughout the globe.

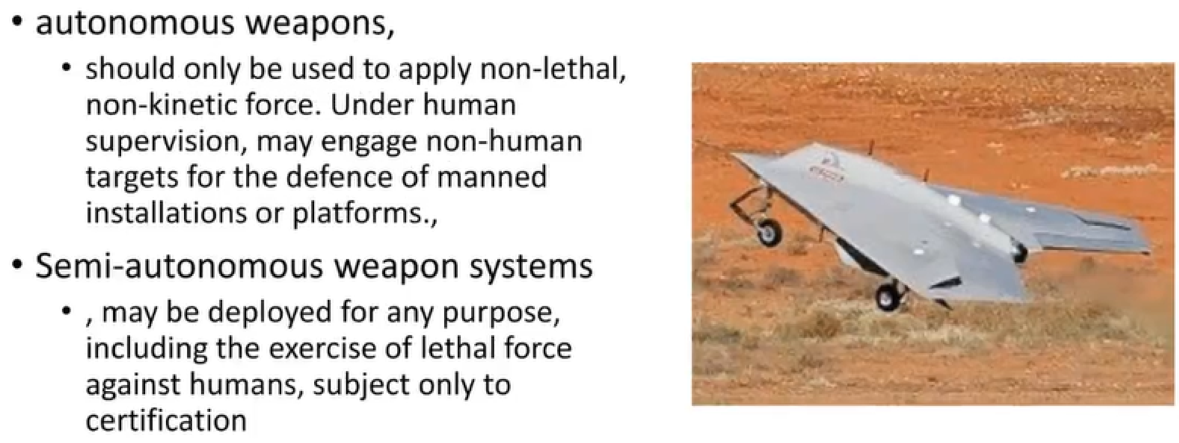


Autonomy in the use of force

- US 2012 Directive on Autonomy in Weapons Systems
  - Target selection, which involves 'the determination that an individual target or a specific group of targets is to be engaged'
  - Autonomous weapons are those 'once activated, can select and engage targets without further intervention by a human operator',
  - Semi-autonomous weapons systems are 'intended to only engage individual targets or specific target groups that have been selected by a human operator'.

This directive defines the difference between autonomous and semi-autonomous devices

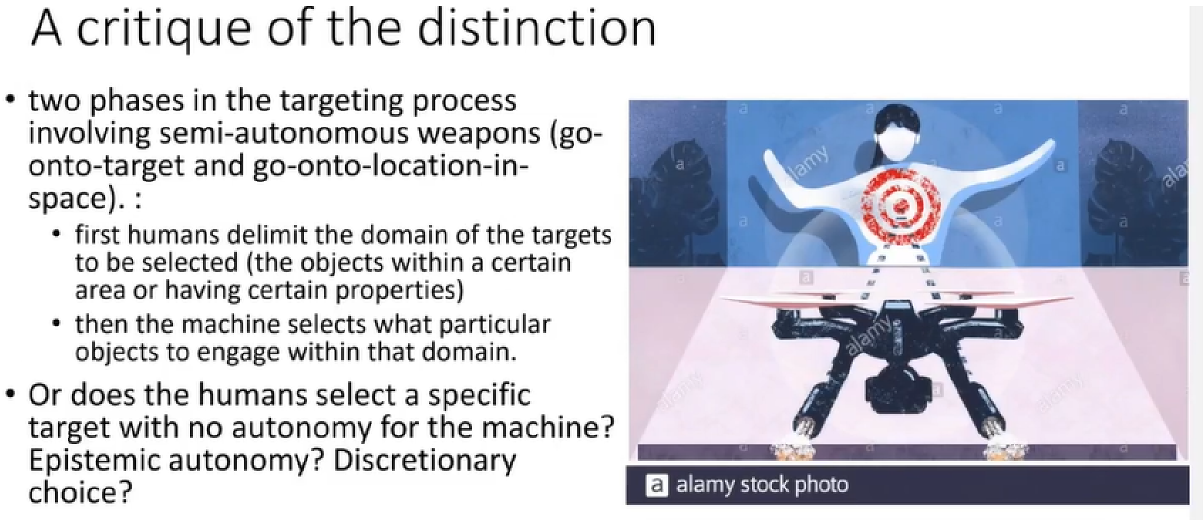
According to this directive, there is no issue in deploying autonomous military vehicles (e.g., tanks, drones) and having a human remotely governing them and hit targets considered to be dangerous after they receive human confermation (*semi-autonomous*). The issue arises when the weapon makes that decision on its own (*autonomous*).



- autonomous weapons,
  - should only be used to apply non-lethal, non-kinetic force. Under human supervision, may engage non-human targets for the defence of manned installations or platforms.,
- Semi-autonomous weapon systems
  - , may be deployed for any purpose, including the exercise of lethal force against humans, subject only to certification

According to the 2012 USA Directive

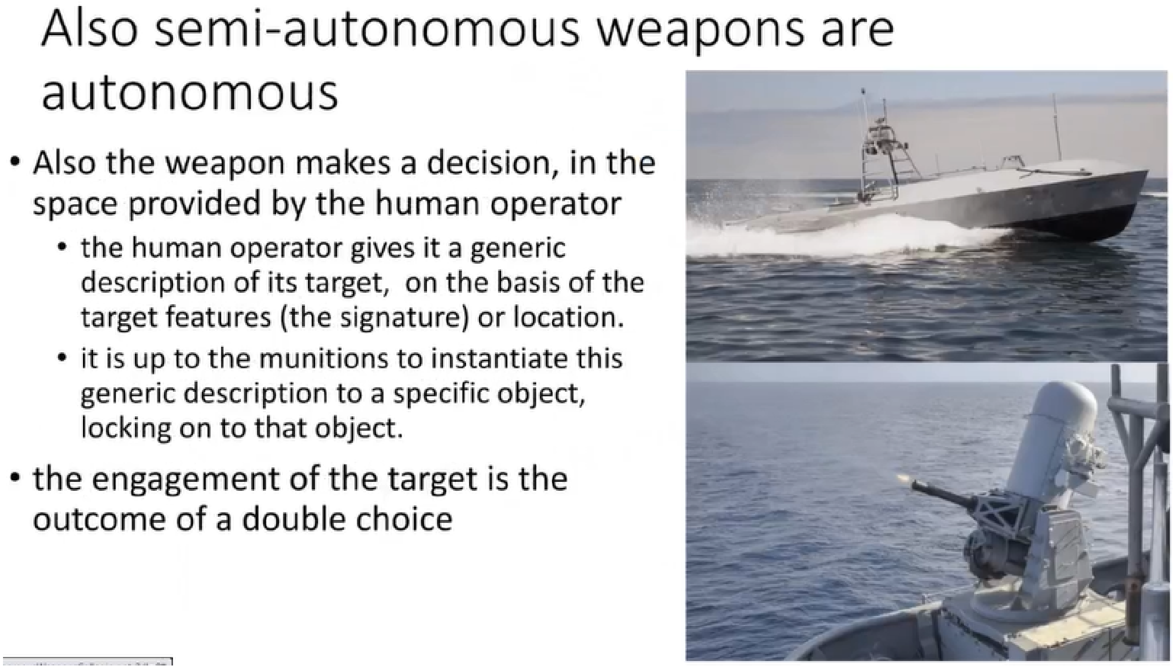
An example of a "lecit" **autonomous weapon** would be the *Iron Dome* of Israel, that prevents enemy missile to hit their targets by intercepting them while they are still flying, neutralizing them before they can cause any harm.



A critique of the distinction

- two phases in the targeting process involving semi-autonomous weapons (go-onto-target and go-onto-location-in-space). :
  - first humans delimit the domain of the targets to be selected (the objects within a certain area or having certain properties)
  - then the machine selects what particular objects to engage within that domain.
- Or does the humans select a specific target with no autonomy for the machine? Epistemic autonomy? Discretionary choice?

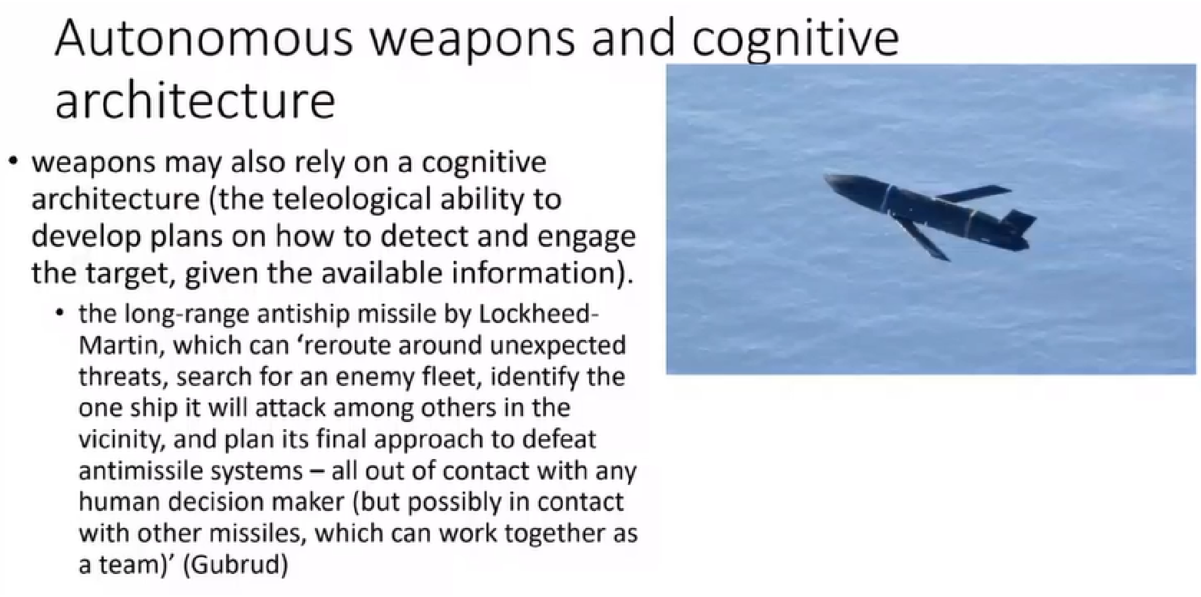
In the two examples above the distinction of the two cases blurs the line between semi and fully autonomous weapons.



Also semi-autonomous weapons are autonomous

- Also the weapon makes a decision, in the space provided by the human operator
  - the human operator gives it a generic description of its target, on the basis of the target features (the signature) or location.
  - it is up to the munitions to instantiate this generic description to a specific object, locking on to that object.
- the engagement of the target is the outcome of a double choice

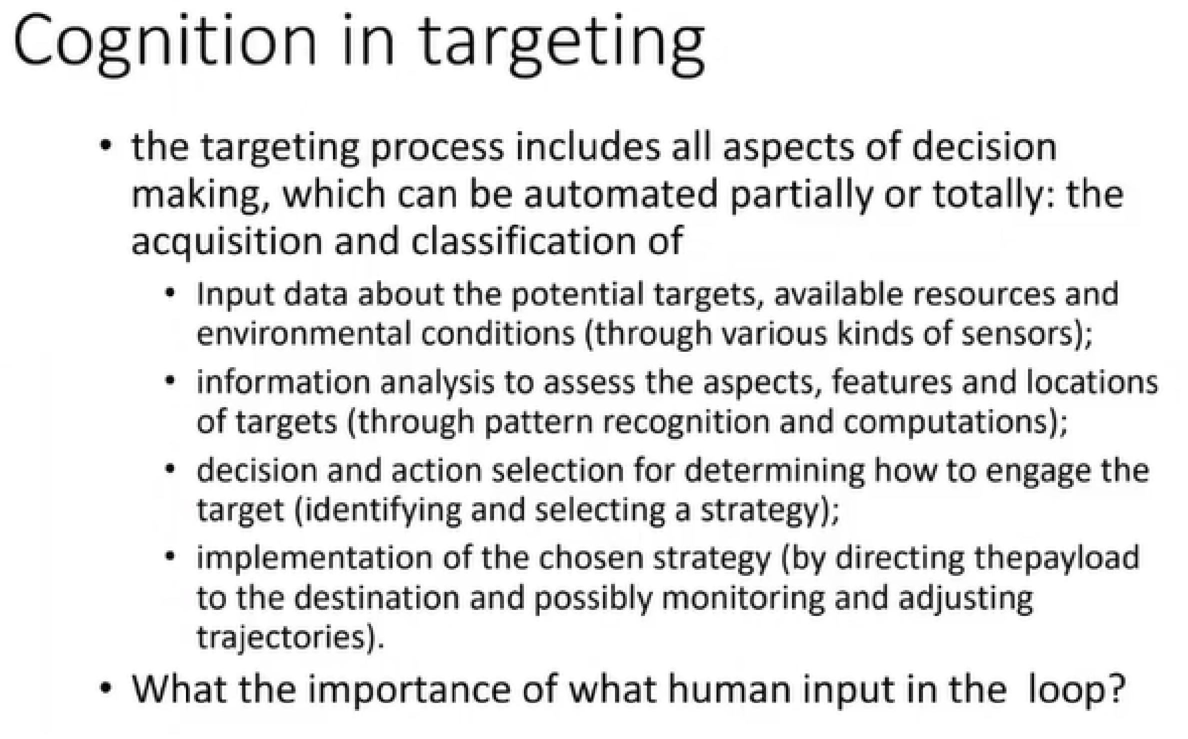
When the weapon has only a generic indications on the target and then it chooses a specific objective, can be still considered as semi-autonomous since it picks the target on its own?



Autonomous weapons and cognitive architecture

- weapons may also rely on a cognitive architecture (the teleological ability to develop plans on how to detect and engage the target, given the available information).
  - the long-range antiship missile by Lockheed-Martin, which can 'reroute around unexpected threats, search for an enemy fleet, identify the one ship it will attack among others in the vicinity, and plan its final approach to defeat antimissile systems – all out of contact with any human decision maker (but possibly in contact with other missiles, which can work together as a team)' (Gubrud)

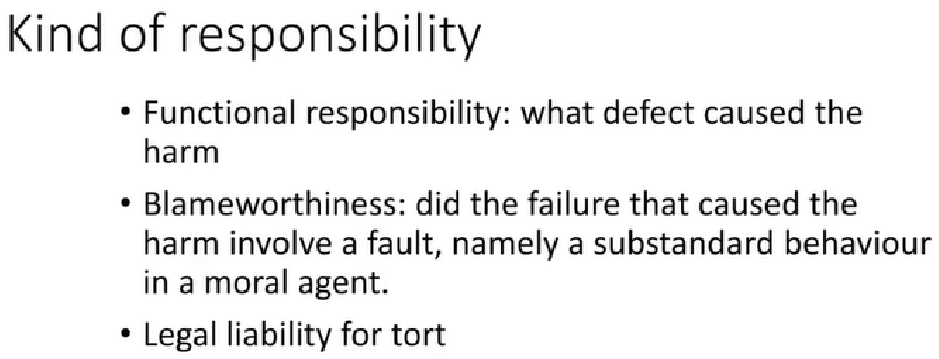
In this case the only human decision made is to launch the missile, otherwise everything else is perfomed by the weapon autonomously.



Cognition in targeting

- the targeting process includes all aspects of decision making, which can be automated partially or totally: the acquisition and classification of
  - Input data about the potential targets, available resources and environmental conditions (through various kinds of sensors);
  - information analysis to assess the aspects, features and locations of targets (through pattern recognition and computations);
  - decision and action selection for determining how to engage the target (identifying and selecting a strategy);
  - implementation of the chosen strategy (by directing thepayload to the destination and possibly monitoring and adjusting trajectories).
- What the importance of what human input in the loop?

When is the targetting process completely autonomous and what is the importance of human input in the loop?



Kind of responsibility

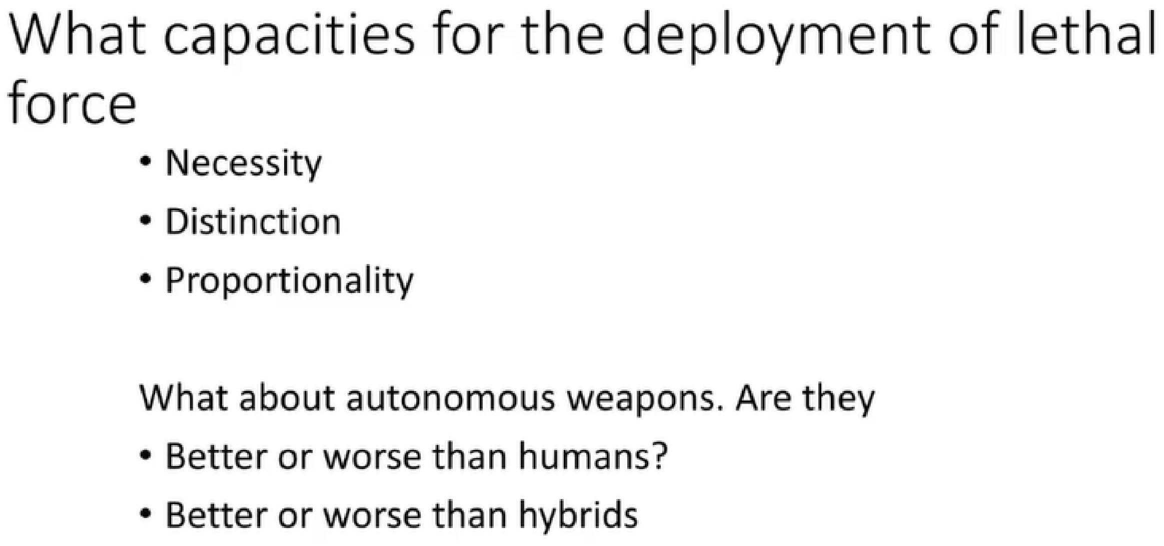
- Functional responsibility: what defect caused the harm
- Blameworthiness: did the failure that caused the harm involve a fault, namely a substandard behaviour in a moral agent.
- Legal liability for tort

The assumption is that there are going to be wars, therefore soldiers and weapons. What would be the role of AI?

There has been a long debate on whether there should be AW or not, what the limitations on development and deployment should be, and in particular if a human should remain in the loop.

It is important to distinguish different kinds of responsibility that may arise from the use of AW.

In particular, laibility: should someone just pay for tort or also there should also be a criminal sanction?



What capacities for the deployment of lethal force

- Necessity
- Distinction
- Proportionality

What about autonomous weapons. Are they

- Better or worse than humans?
- Better or worse than hybrids

Aspects of the Laws of war:

- ***Jus ad bellum:*** when it is justified to enter the war with another country; the treaty of the UN forbids aggressive wars and allows only the defensive ones. (Humanitarian wars challenge this concept)

- ***Jus in bellum:*** how we should behave when we are at war. This defines what is a legitimate military action in war. This is also called *Humanitarian Law.*

Three basic principles derive from correct behaviour in war:

- ***Necessity:*** when harm is caused to the other party, it must be justified by the purpose that we want to achieve otherwise it would be unmotivated cruelty.
- ***Distinction***: all military activities should be against the enemy army and not the civilians, if they are harmed it must be due to side effects. Bombardments against the german population during WWII violated this principle.
- ***Proportionality:*** the harm caused to the civilians must be proportional to the military goal pursued.

What effects can imply the introduction of AW? Would they respect those principles more than humans?

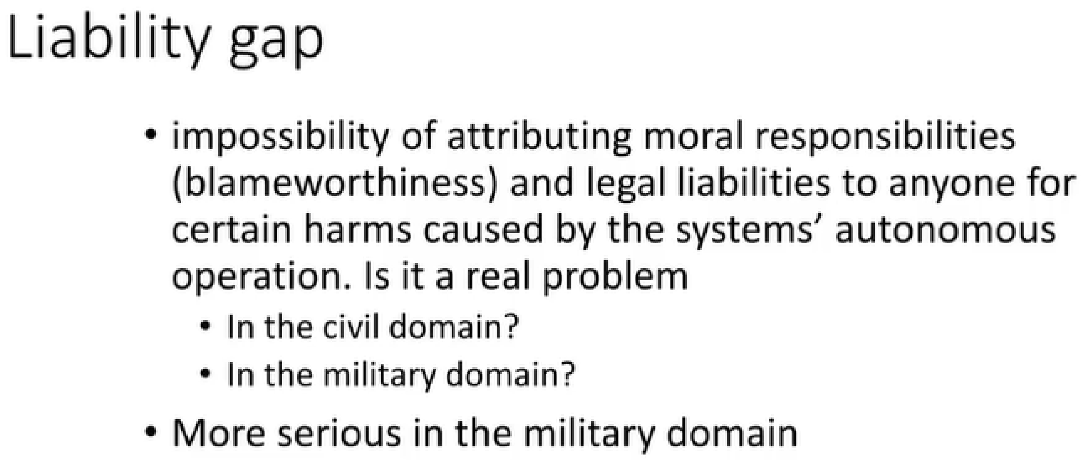
There have been episodes of harm to the prisoners taken during a military action, which violates the humanitarian law, that might have been due to an emotional reaction of human soldiers, something that AW would not be able to do.

On the other hand, there are some critisms that an AW would not be able to distinguish military forces from civilians. Moreover, it wouldn't be able to assess the importance of a certain goal w.r.t. civilians live (principle of proportionality). Also, the development of AWs would give life to another arms race to prevent possible enemies to be more advanced (like it happened in the past).

In the past there have been effective bans on certain kinds of weapons, gasses for examples, which are easily identifiable. In the case of AWs, the technological advancements would make it harder to isolate a specific tipology of technology, since they can be used in different purposes where they would not cause harm (e.g, face recognition systems) (also military vs civilian deployment).

It has been argued that AWs could induce more advanced countries to engage in war activities more easily, as they could deploy their technology instead of their soldiers, whose lives would be safe. War activities would be more attractive.

According to Sartor, it is impossible to fully exclude the use of AI in a military context, as there are powerful incentives to keep using it (many countries already do that, e.g. Israel and USA). What we have to ensure is that humans have the final decision when they engage in lethal actions.



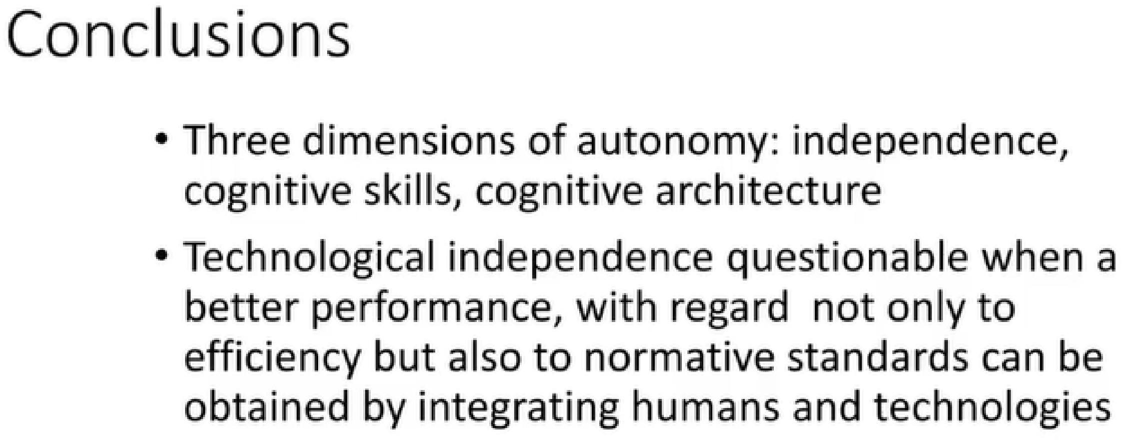
Liability gap

- impossibility of attributing moral responsibilities (blameworthiness) and legal liabilities to anyone for certain harms caused by the systems’ autonomous operation. Is it a real problem
  - In the civil domain?
  - In the military domain?
- More serious in the military domain

Who should be held responsible if the AW kills the wrong target or there are collateral damages after that the human have confirmed the lethal action?

According to Sartor: it is difficult to identify individuals that are clearly responsible in this case, but we must consider that rarely military have been called upon to respond for their war actions to begin with.

The liability gap would not be the major issue in this domain, and the respect of the humanitarian law should be the priority and also to avoid an arms race to develop more destructive weapons. So, to prevent major damages AWs should be banned and there should be an international treaty that regulates the situation.



Conclusions

- Three dimensions of autonomy: independence, cognitive skills, cognitive architecture
- Technological independence questionable when a better performance, with regard not only to efficiency but also to normative standards can be obtained by integrating humans and technologies