Autonomous Weapons Systems and International Law

A Case Study on Human-Machine Interactions in Ethically and Legally Sensitive Domains

Prof. Daniele Amoroso

Università degli studi di Cagliari

A look at the bigger picture: the Fourth Industrial Revolution

The Fourth Industrial Revolution builds on the "staggering confluence of emerging technology breakthroughs, covering wide-ranging fields such as artificial intelligence (AI), robotics, the internet of things (IoT), autonomous vehicles, 3D printing, nanotechnology, biotechnology, materials science, energy storage and quantum computing"

(Klaus Schwab, World Economic Forum, 2016)

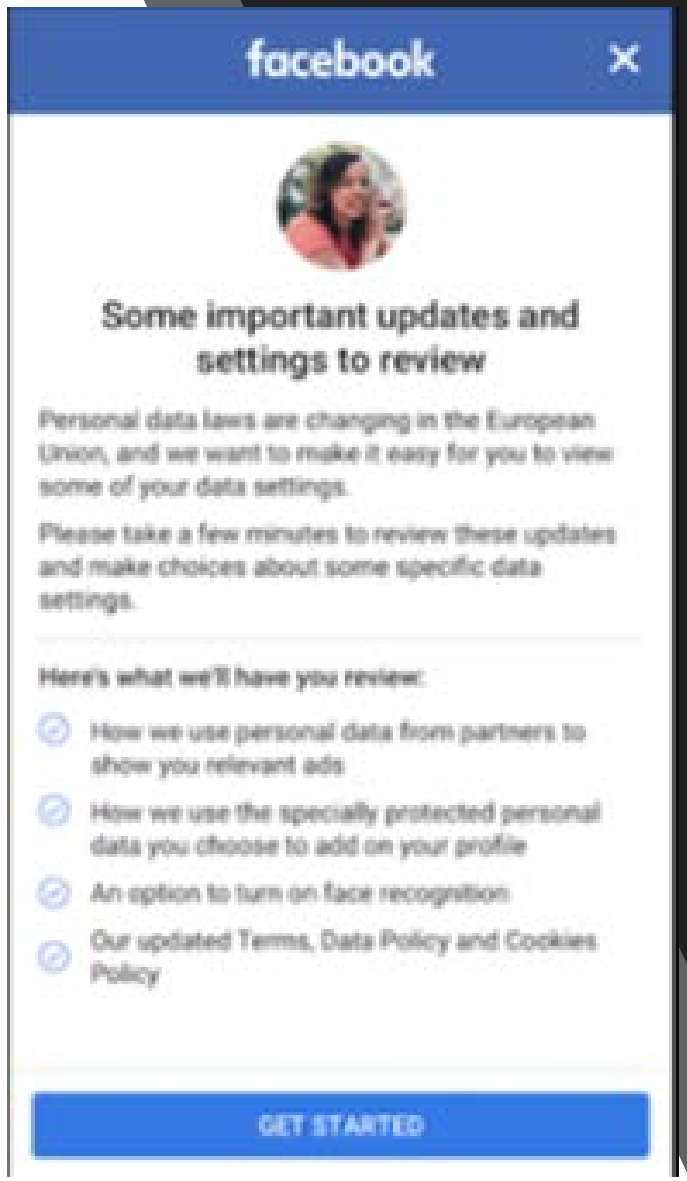
What implications for the society at large?

Autonomous data processing to *support* human decision-making

The Public Safety Assessment (PSA) software

- The pre-trial judge must decide whether the defendant should remain in jail while awaiting trial or can be safely released until the court date
- The PSA algorithm, used by several State courts in the US, predicts whether a defendant will commit a new crime or fail to return to court on the basis of nine factors (including seriousness of the crime, age, prior conduct of the defendant)
- Humanitarian purposes:
  - To avoid that pre-trial release is based solely on the possibility to pay the bail (money = freedom)
  - To avoid human biases leading to discrimination on the grounds of race, gender, employment status, level of education, or history of substance use

COMPAS recidivism black bias



Autonomous data processing to replace human decision-making

The EU General Data Protection Regulation (EU) 2016/679

- The problem of "profiling" (Recital 71)

  [A]ny form of automated processing of personal data evaluating the personal aspects relating to a natural person [… that] produces legal effects concerning the [data subject] or similarly significantly affects him or her (Recital 71)

- The right to a human decision (Article 22)

  The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her

When AI meets Robotics: Autonomous Vehicles

Main ethical-legal issues

- Road safety as a "push factor" for more autonomy
- The “inevitable accidents” problem
  - Should the AV minimize the harm to bystanders or save the life of the occupants? Who has to decide?
- Responsibility for harmful events

Overarching issues

- *Technical question*: is it possible to design an artificial agent that is able to autonomously comply with legal prescriptions, especially when the latter require the application of discretionary reasoning and/or equitable evaluations?
- *(Strictly) legal question*: how to allocate responsibility in case of harmful events caused by the machine?
- *Ethical (deontological) question*: is it morally acceptable to remove human agency from decision-making processes that are likely to impinge on individual rights?
- *Ethical (consequentialist) question*: is it morally required to replace human operators with autonomous machines when the latter's performances ensure a better protection of the interests at stake?

They are all reflected in the debate on Autonomous Weapons Systems!

The origins of the debate on AWS: A tale of two roboticists

"[R]obots not only can be better than soldiers in conducting warfare in certain circumstances, but they also can be more humane in the battlefield than humans"

Ronald Arkin

"Rather than making war more humane and ethical, autonomous armed robotic machines are simply a step too far in the dehumanization of warfare. We must continue to ensure that humans make the moral decisions and maintain direct control of lethal force"

Noel Sharkey

ICRC

Convention on Certain Conventional Weapons (CCW)
Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)
11-15 April 2016, Geneva

1. Definitions

The ICRC has defined autonomous weapon systems as: "Any weapon system with autonomy in its critical functions. That is, a weapon system that can select (i.e. search for or detect, identify, track, select) and attack (i.e. use force against, neutralize, damage or destroy) targets without human intervention."

Department of Defense

DIRECTIVE

NUMBER 3000.09
November 21, 2012
Incorporating Change 1, May 8, 2017

USD(P)

SUBJECT: Autonomy in Weapon Systems

autonomous weapon system. A weapon system that, once activated, can select and engage targets without further intervention by a human operator.

CAMPAIGN TO STOP KILLER ROBOTS

Urgent Action Needed to Ban Fully Autonomous Weapons
Non-governmental organizations convene to launch Campaign to Stop Killer Robots

(London, April 23, 2013) – Urgent action is needed to pre-emptively ban lethal robot weapons that would be able to select and attack targets without any human intervention, said a new campaign launched in London today. The Campaign to Stop Killer Robots is a coordinated international coalition of non-governmental organizations concerned with the implications of fully autonomous weapons, also called "killer robots."

WHAT IS AN "AUTONOMOUS WEAPONS SYSTEM"?

A weapon system that – by relying on (more or less) advanced AI technologies – is able, once activated, to select and engage targets without human intervention

Existing AWS

Air defensive systems

Nächstbereichschutzsystem MANTIS – Germany

Sentry Robots

Super aEgis II – South Korea

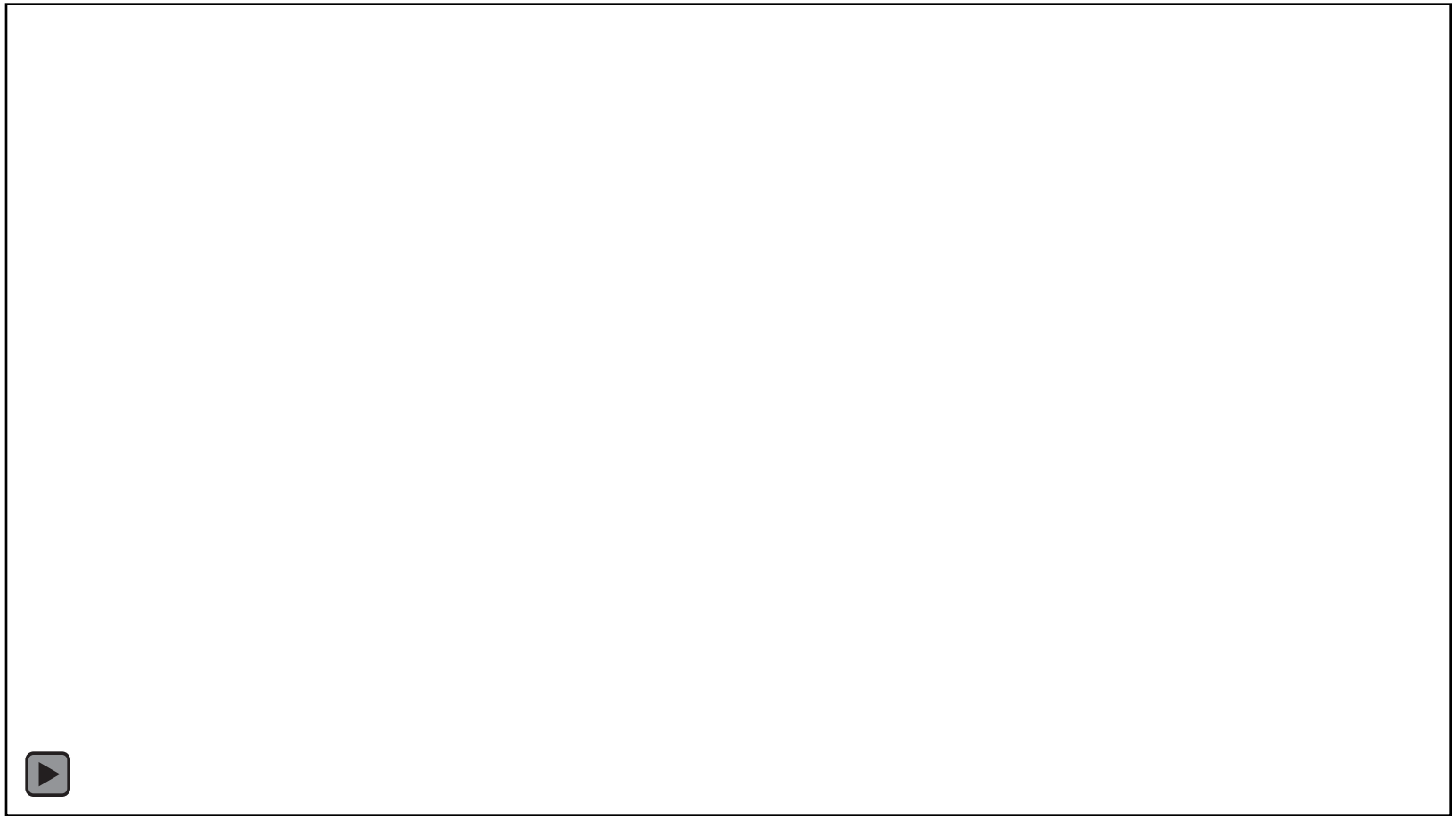
Fire-and-forget munitions

Brimstone – UK

Loitering munitions

Harpy NG – Israel

What about the future?

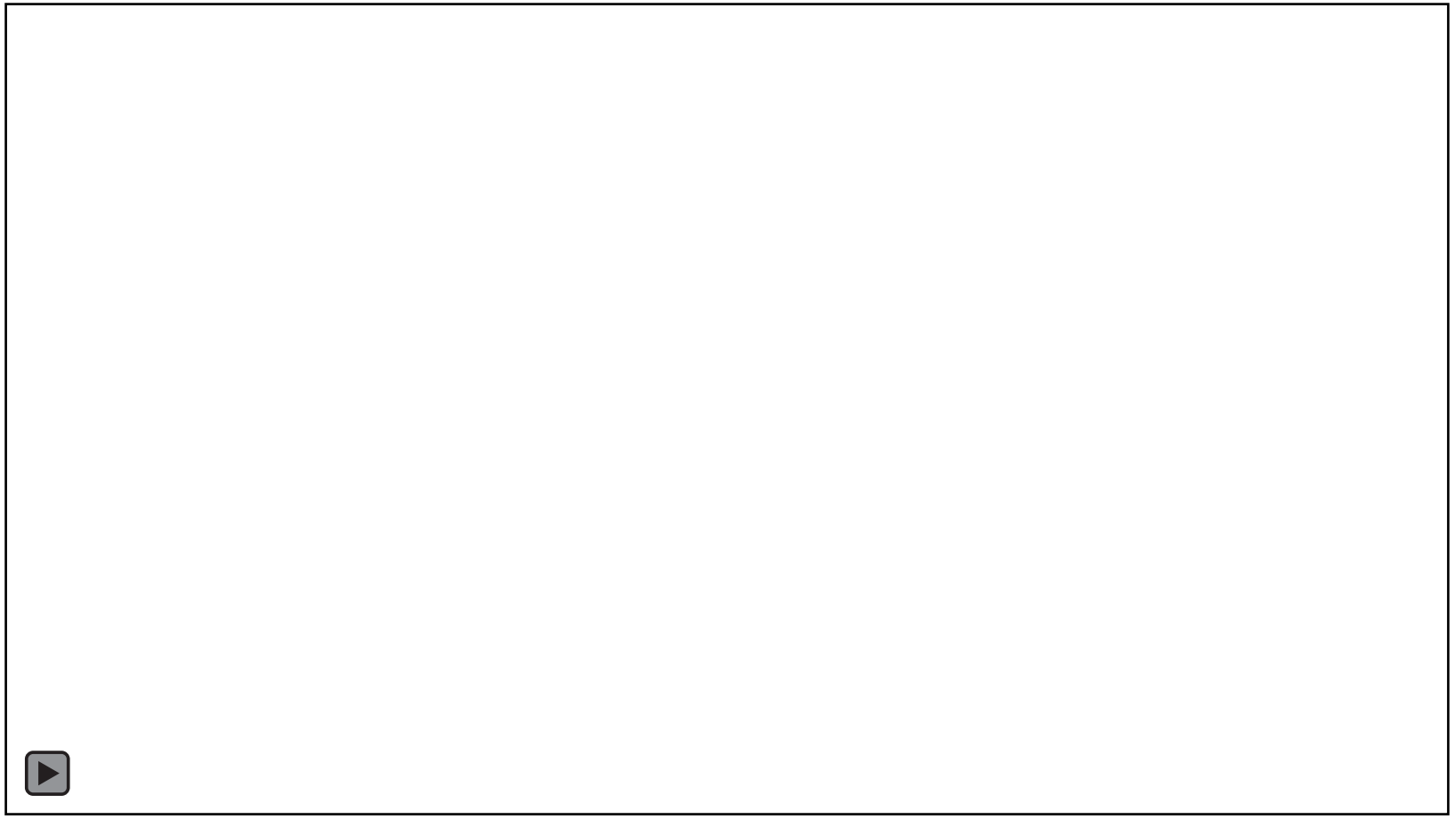


…is it just science fiction?

Swarms of micro-drones
Perdix Project – US

*"Due to the complex nature of combat, **Perdix** are not pre-programmed synchronized individuals, they are a collective organism, sharing one distributed brain for decision-making"*

William Roper
Director of the US Strategic Capabilities Office
(9 January 2017)



WHAT IS "SPECIAL" IN THE FUNCTIONS OF SELECTING AND ENGAGING TARGET?

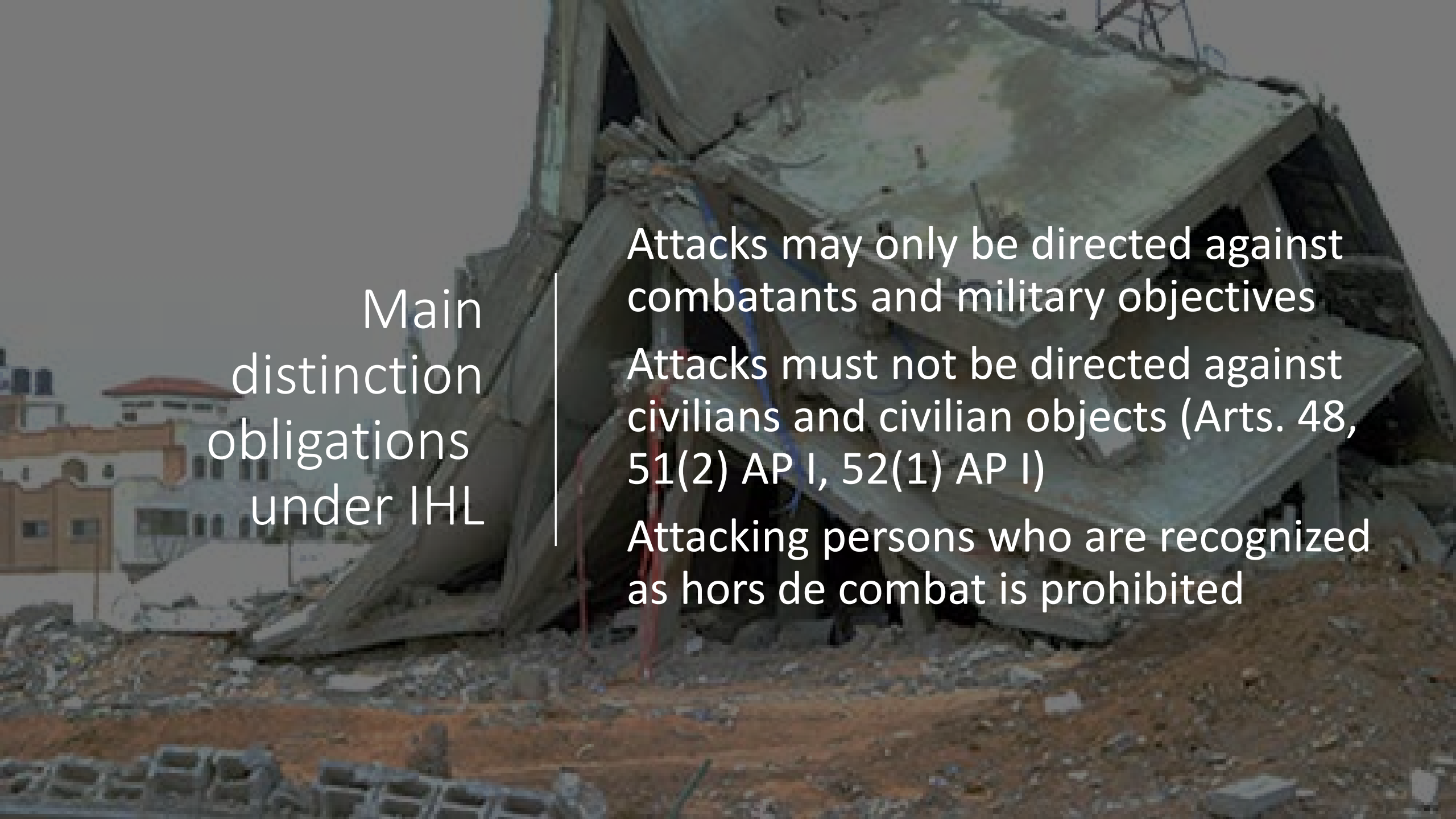
They are crucially regulated by international humanitarian law (IHL)

Their performance is a key factor for the purposes of responsibility ascription

They imply moral choices that affect legally protected individual positions

The issues at stake in the debate on AWS

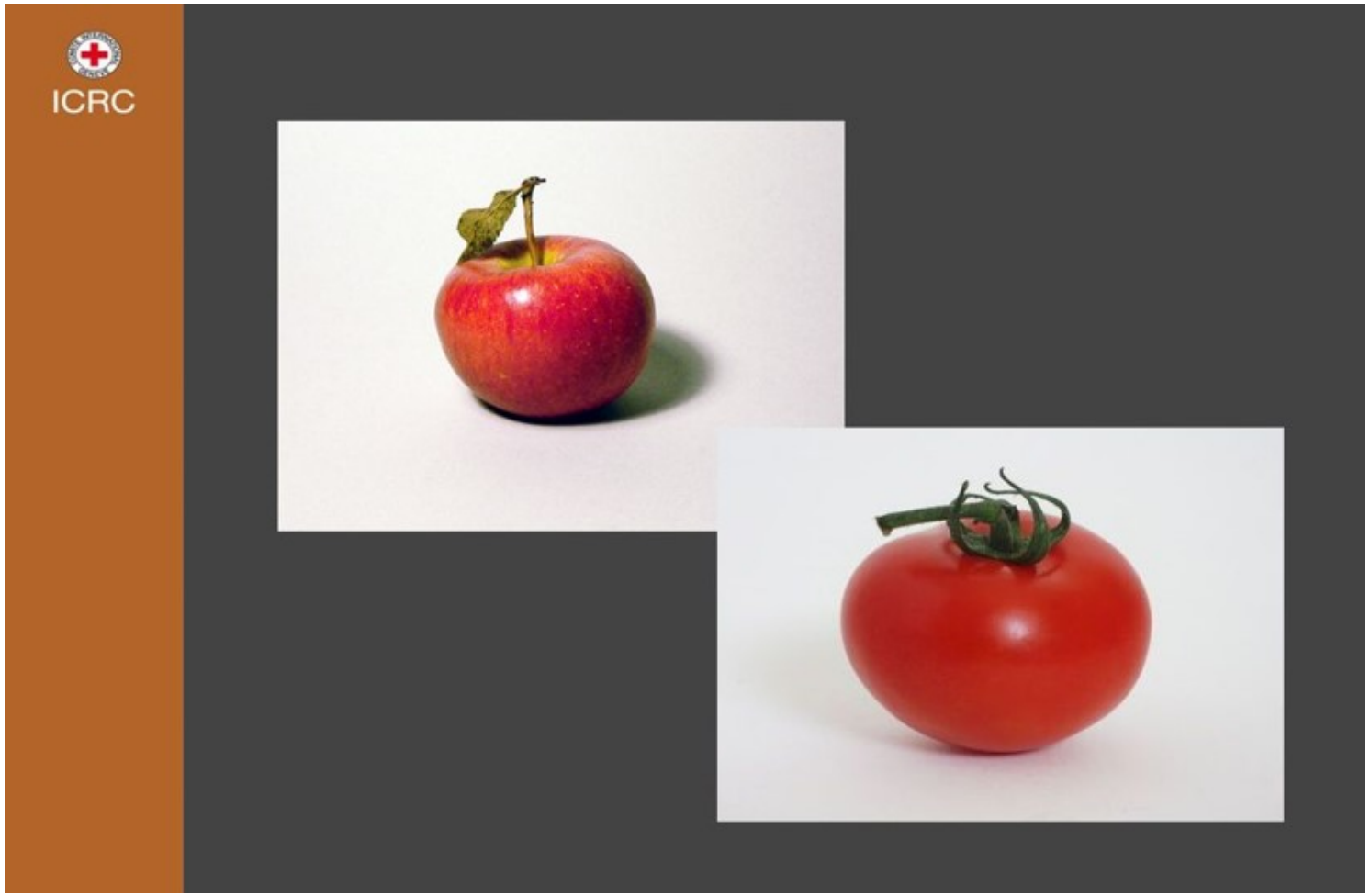
- Will AWS ever be able to ensure proper compliance with International Humanitarian Law (IHL)?
  - Will AWS ever be able to distinguish military objectives from protected persons and objects?
  - Can AWS be programmed to perform proportionality assessments?
- Who is to be held responsible in case an AWS takes unlawful targeting decisions?
- Is it ethically and legally acceptable to delegate lethal decision-making to an artificial agent?



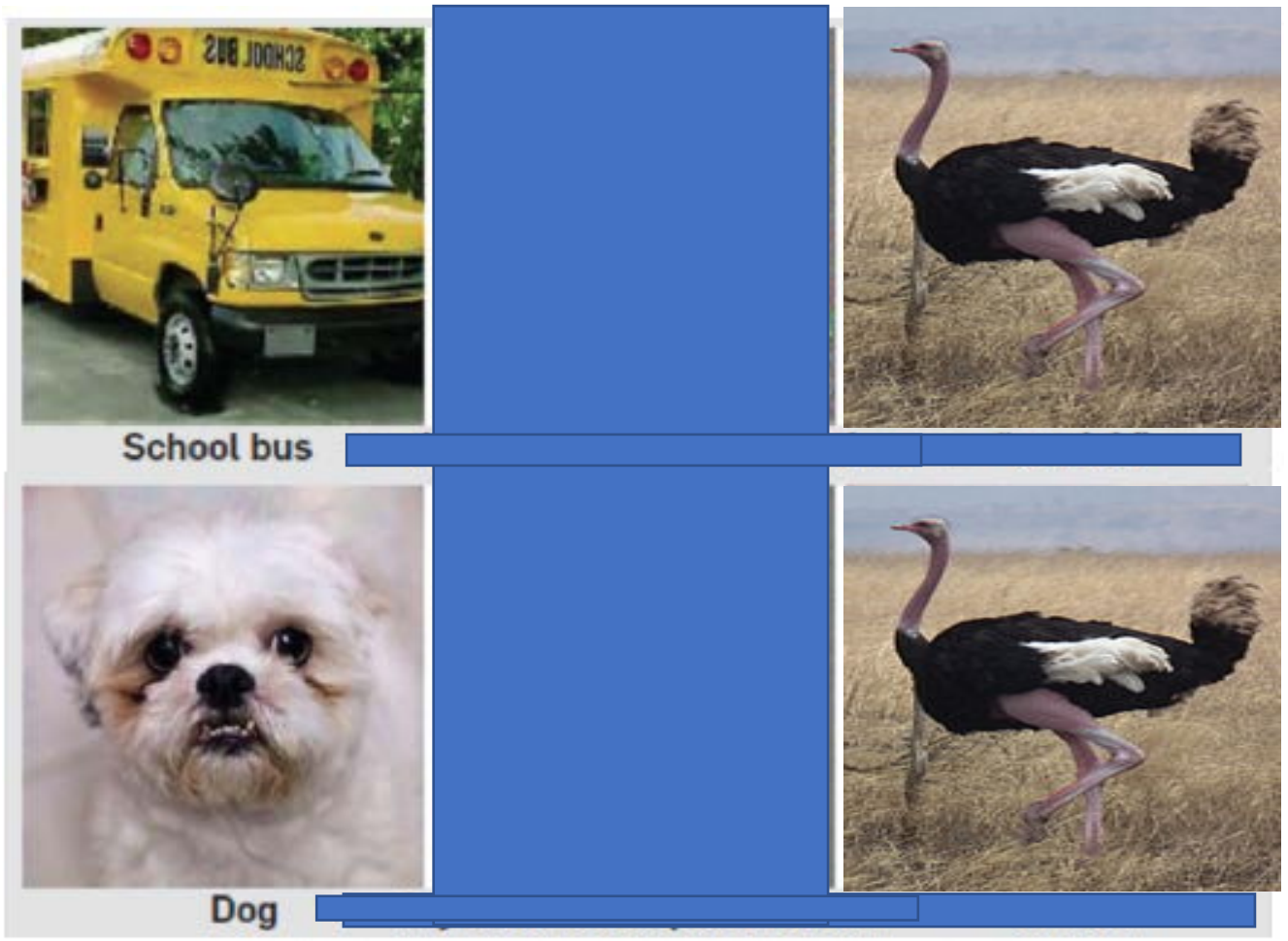
Main distinction obligations under IHL

- Attacks may only be directed against combatants and military objectives
- Attacks must not be directed against civilians and civilian objects (Arts. 48, 51(2) AP I, 52(1) AP I)
- Attacking persons who are recognized as hors de combat is prohibited

Problems with autonomous image recognition



Autonomous image perturbation

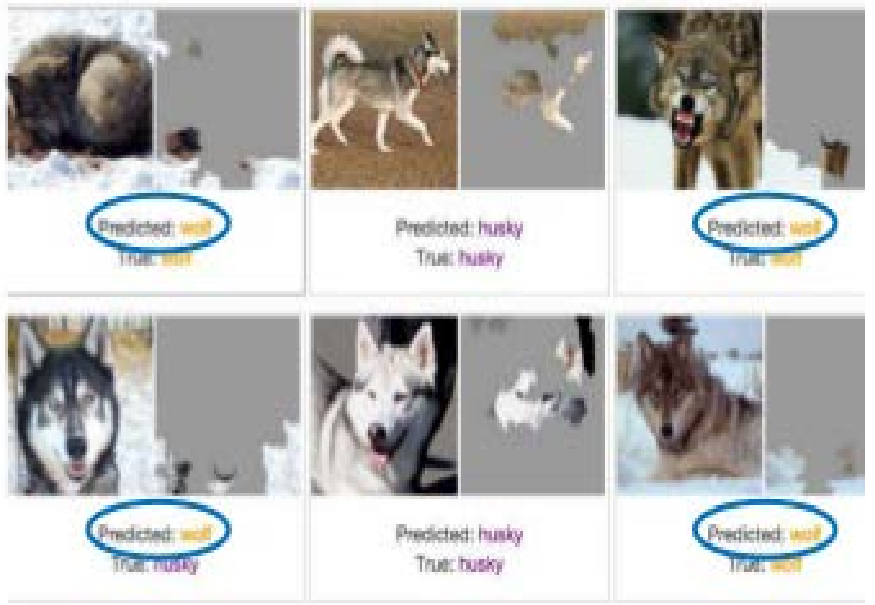


Szegedi C. et al. (2014). Intriguing properties of Neural Networks

The "semantic gap" between humans and machines

Even when machines prove better than humans in a given task, they remain subject to serious and counter-intuitive mistakes, which a human would have never made

This is because only humans have a proper understanding of the meaning or concept of the objects they classify and recognize (semantic gap)



(b) (c) (d)

Issues of distinction are not limited to object perception!

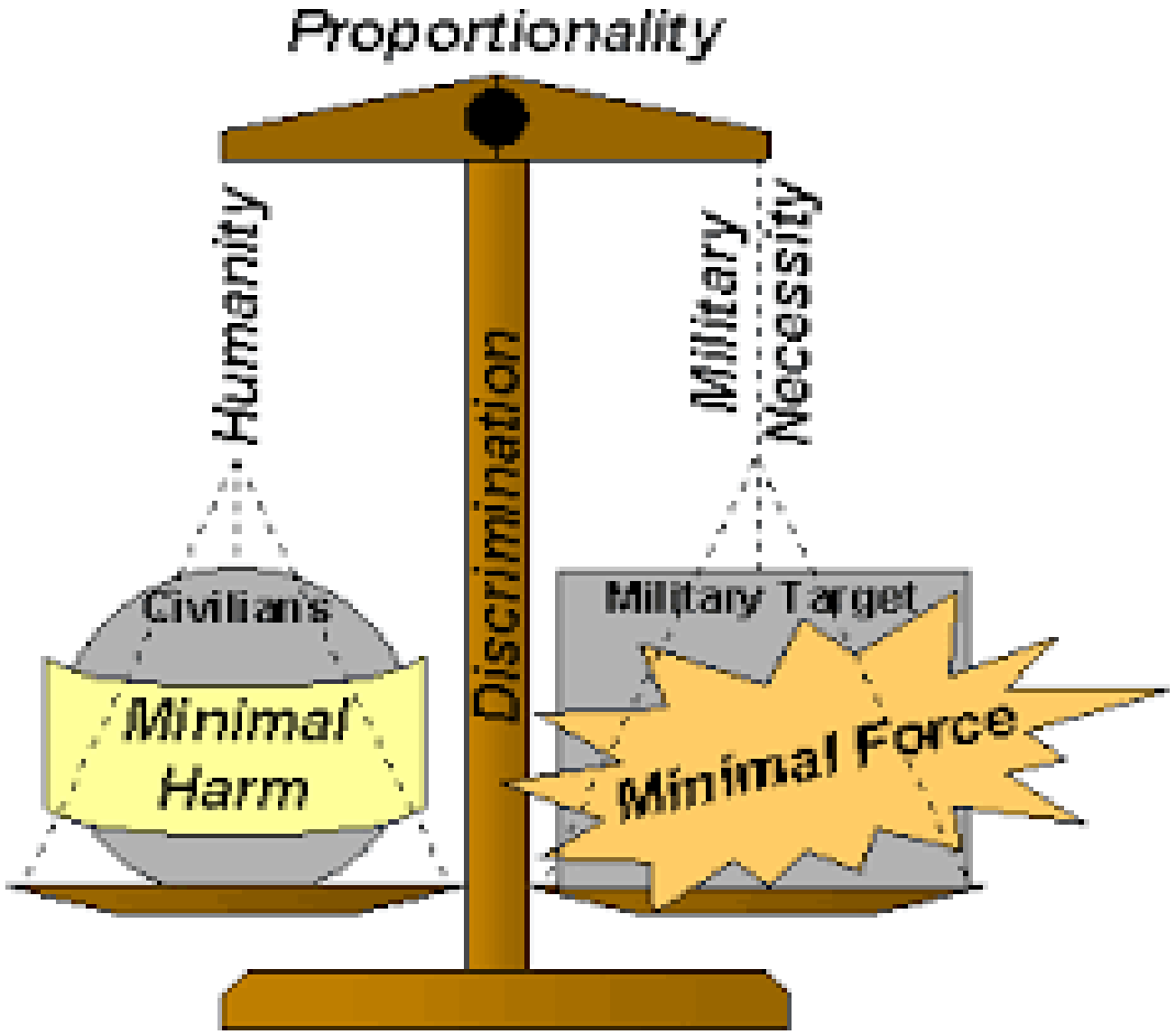
Article 41 of the First Additional Protocol to the Geneva Conventions

1. A person who is recognized or who, in the circumstances, should be recognized to be *hors de combat* shall not be made the object of attack.

2. A person is *hors de combat* if:
[…] (b) he clearly expresses an intention to surrender; or
(c) he has been rendered unconscious or is otherwise incapacitated by wounds or sickness, and therefore is incapable of defending himself;
provided that in any of these cases he abstains from any hostile act and does not attempt to escape.

The principle of proportionality in IHL

It is prohibited to launch an attack that may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated



Is it possible to translate the principle of proportionality into a computer code?

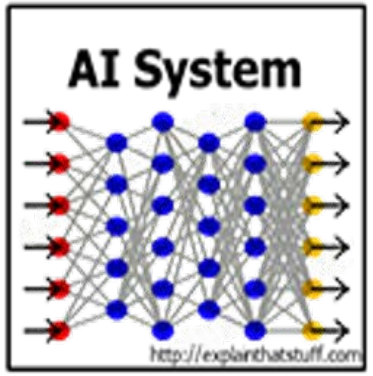
"[W]eighing the expected collateral damage against the anticipated military advantage will never be a job for one's pocket calculator" (Kalshoven, 1992)

Why?

The principle of proportionality "applies to situations in which a number of different interests entitled to legal protection can be identified but cannot be a priori composed in a predetermined behavioural scheme" (Cannizzaro, 2014)

The principle of proportionality, therefore, cannot be pre-programmed because it does not provide "a 'reference answer' which can be considered to solve the moral dilemma in a generally satisfactory way" (Matthias, 2011)

The "accountability gap" problem



- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

- AWS are going to be more and more unpredictable because of:
  - Machine learning architectures (Matthias 2004)
  - Features of the operational environment
- As a consequence, it will be extremely difficult (when not impossible) to establish a normatively acceptable responsibility ascription for lack of *mens rea*

A RADICAL ETHICAL OBJECTION AGAINST AUTONOMY IN WEAPONS SYSTEMS

- Delegation of lethal decision-making to an artificial agent is ethically inacceptable
  - this claim is legally relevant under the principle of human dignity

Two argumentative variants

- Targeted people have the right not to be subject to automated decision-making interfering with their life or physical integrity
- Human beings, as moral agents, should bear the burden of taking lethal decisions, irrespective of their (un)lawfulness

A POSSIBLE WAY OUT: THE NOTION OF MEANINGFUL HUMAN CONTROL (MHC)

- Growing international consensus on the idea that every weapons system should be subject to a meaningful human control (MHC)
- Threefold role for human control to be "meaningful"
  - to prevent a malfunctioning of the weapon from resulting in a direct attack against the civilian population, or in excessive collateral damages (fail-safe mechanism)
  - to secure the legal conditions for responsibility ascription in case a weapon follows a course of action that is in breach of international law (accountability attractor)
  - to ensure that decisions affecting the life, physical integrity and property of people involved in armed conflicts are not taken by artificial agents (moral agency enactor)

ELEMENTS OF MEANINGFUL HUMAN CONTROL

- Primary obligations
  - Exclusive control privileges for human operators (e.g. approval of target selection, veto power, etc.)
  - Need for a differentiated approach (*one size does not fit all*)
    - … limited autonomy may be allowed under certain conditions (e.g. for defensive systems)
- Ancillary obligations
  - Crucial to ensure effective human-weapon interaction
  - Training
  - Design: interpretability and explainability

«TRANSPLANT» OF THE NOTION OF MHC IN OTHER DOMAINS

- Surgical robots (Ficuciello et al, 2019)
- Autonomous vehicles (Santoni de Sio et al, 2019)
- Judicial applications of AI (Amoroso and Tamburrini, 2020)

ANY QUESTIONS/COMMENTS?