AI and Warfare:
Navigating the Ethical Frontlines

Raja Chatila
Institute of Intelligent Systems and Robotics (ISIR)
Faculty of Sciences and Engineering
Sorbonne University, Paris

Raja.Chatila@sorbonne-universite.fr

International Humanitarian Law

- Foundation: *limiting the use of violence in armed conflicts by sparing those who do not or no longer directly participate in hostilities*
- Principles:
    - the principle of **humanity** (the “elementary” considerations of humanity being reflected and expressed in the Martens clause)
    - the principle of **distinction** between civilians and combatants, and between civilian objects and military objectives
    - the principle of **proportionality** prohibiting attacks “which 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”
    - the principle of **military necessity** (from which flows the prohibition of superfluous injury and unnecessary suffering).

Source ICRC: https://casebook.icrc.org/a_to_z/glossary/fundamental-principles-ihl

The Martens Clause (1899)

“Until a more complete code of the laws of war has been issued, the High Contracting Parties deem it expedient to declare that, in cases not included in the Regulations adopted by them, the inhabitants and the belligerents remain under the protection and the rule of the law of nations, as they result from the usages established among civilized peoples, from the **laws of humanity** and the dictates of public **conscience**.”

Article 36 - New weapons

Geneva Convention 1949 - Additional Protocol 1977

In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party.

How it started at the UN CCW

- 2013: Following a report by UN Special Rapporteur on extrajudicial, summary or arbitrary executions Christof Heyns, the CCW Meeting of High Contracting Parties (HCP) decided that the Chairperson will convene in 2014 an Informal Meeting of Experts to discuss the questions related to emerging technologies in the area of lethal autonomous weapons systems (LAWS).
- 2014: The first Informal Meeting of Experts is held in accordance with the decision of the 2013 CCW Meeting of HCPs. Chair: Ambassador Simon-Michel of France
- 2015 & 2016 : The second and third Informal Meeting of Experts is held. Chair: Ambassador Michael Biontino of Germany
- 2016: At the CCW Fifth Review Conference, HCPs decide to establish an open-ended Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems. Mandate: to build on the work of the previous meetings of experts and to explore and agree on possible recommendations on options related to emerging technologies in the area of LAWS, in the context of the objectives and purposes of the Convention.
- …
- 2018: The Group of Governmental Experts meets for 10 days and affirms 10 guiding principles. Chair: Ambassador Amandeep Singh Gill of India
- …

What are LAWS?

*“At present, no commonly agreed definition of Lethal Autonomous Weapon Systems (LAWS) exists.”**

- **ICRC**: *AWS are weapons that, once activated, can identify, select and apply force to targets without human intervention.*

- *Any weapon system with **autonomy in its critical functions**—that is, a weapon system that can (search for, detect, identify, track or select) and attack (use force against, neutralize, damage or destroy) targets without human intervention.*

- Instead of a precise target, the AW must be provided with a description (a signature) characterizing the targets, that can be recognized by the machine, and a spatial/temporal region of operation.

*https://disarmament.unoda.org/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/

Do AWS exist?

- **Germany**: *LAWS [are] weapons systems that completely exclude the human factor from decisions about their employment. Emerging technologies in the area of LAWS need to be conceptually distinguished from LAWS. Whereas emerging technologies such as digitalization, artificial intelligence and autonomy are integral elements of LAWS, they can be employed in full compliance with international law.*
- **France:**
  - **“fully” lethal autonomous weapons systems:** systems capable of acting without any form of human supervision or dependence on a command chain by setting their own objectives or by modifying, without any human validation, their initial programme or their mission framework)
  - **“partially” autonomous lethal weapons systems:** lethal weapons systems featuring decision-making autonomy in critical functions such as identification, classification, interception and engagement to which, after assessing the situation and under their responsibility, the military command can assign the computation and execution of tasks related to critical functions within a specific framework of action

Do AWS exist?

Russia: There is no consensus definition of LAWS in existing international law. Since the issue pertains to prospective types of weapons, the definition of LAWS should not be interpreted as limiting technological progress and detrimental to research on peaceful robotics and artificial intelligence. The definition of LAWS should meet the following requirements:

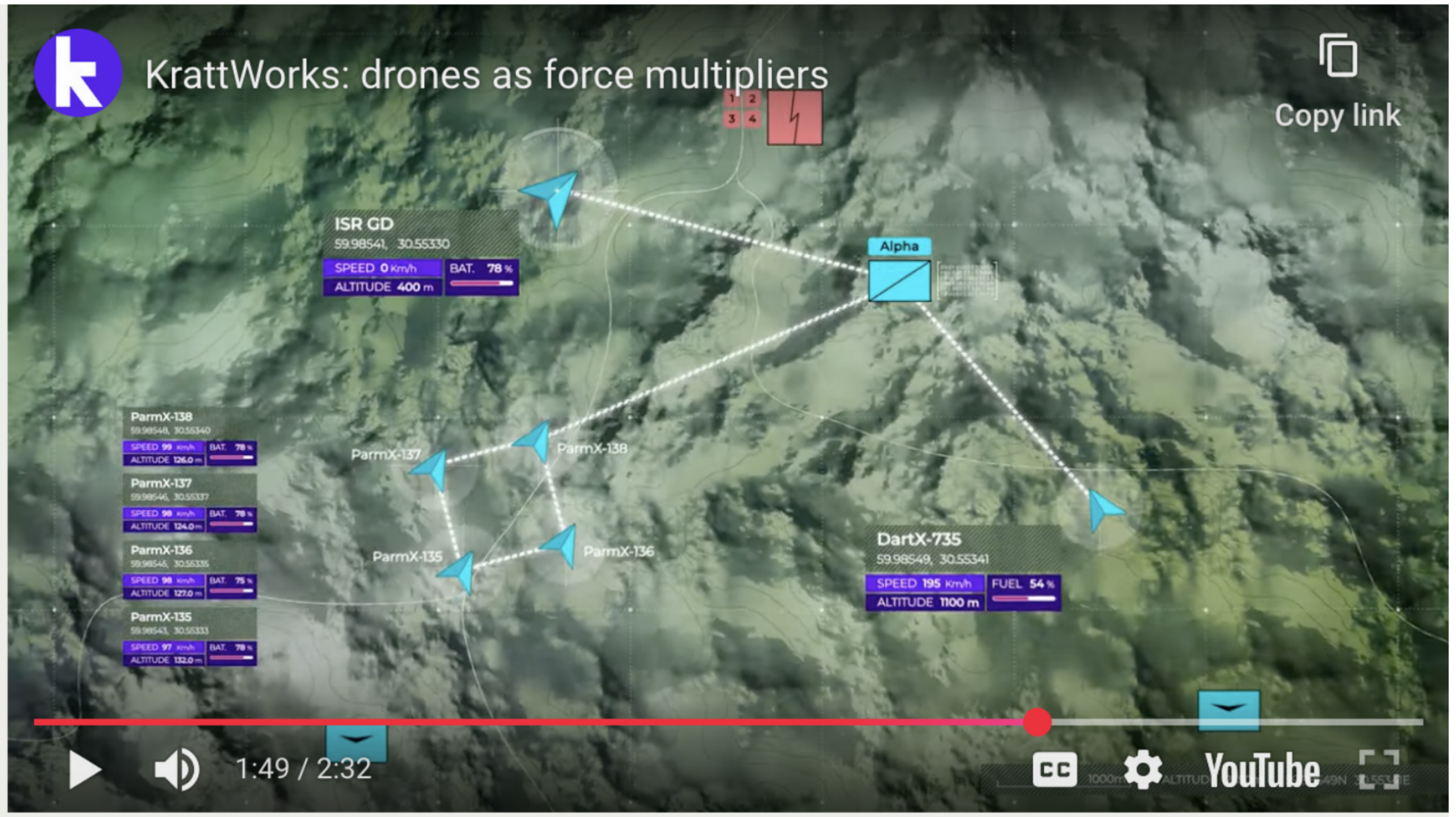
  - contain the description of the types of weapons that fall under the category of LAWS, conditions for their production and testing as well as their usage procedure;
  - not be limited to the current understanding of LAWS, but also take into consideration the prospects for their future development;
  - be universal in terms of the understanding by the expert community comprising scientists, engineers, technicians, military personnel, lawyers and ethicists.

- A lethal autonomous weapons system is a fully autonomous unmanned technical means other than ordnance that is intended for carrying out combat and support missions without any involvement of the operator.

Do AWS exist?

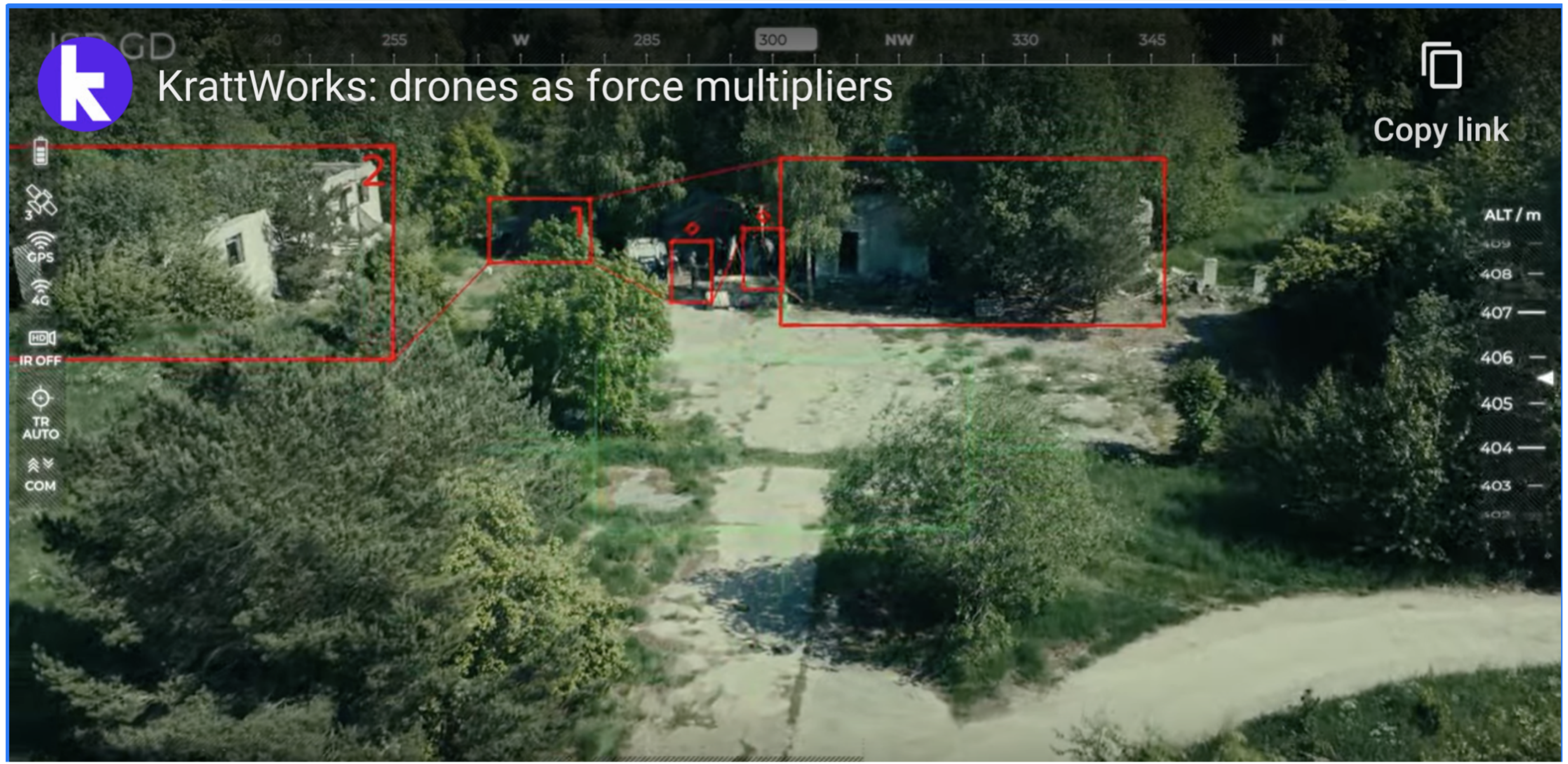
Australia, Canada, Japan, S. Korea, UK, USA:

- …. Recognizing that the research and development of new technologies in the field of artificial intelligence is progressing at a rapid pace, potentially enabling novel and more sophisticated weapons with autonomous functions, including those weapon systems that, once activated, can identify, select, and engage targets with lethal force without further intervention by an operator (“autonomous weapon systems”) for the purposes of these draft articles and without prejudice to any other understandings of this or similar terms for other purposes.



A promotional video from Krattworks depicts scenarios in which the company's drones augment soldiers on offensive maneuvers. KRATTWORKS

Asia Times



IEEE Spectrum, June 2, 2025

AI System Definition

In the EU AI Act - similar to the OECD's

- "An AI system is a machine-based system that, for **explicit or implicit objectives**, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or **decisions** that can influence physical or virtual **environments**. Different AI systems vary in their levels of **autonomy** and adaptiveness after deployment."

a. Build phase, pre-deployment
ENVIRONMENT
(CONTEXT)
physical
/ virtual
Perceive
Objectives
explicit /
implicit
AI SYSTEM
DATA & INPUT
AI MODEL
1. BUILD PHASE
OUTPUTS
b. Use phase, post-deployment
ENVIRONMENT
(CONTEXT)
physical
/ virtual
Perceive
Objectives
explicit /
implicit
Influence
AI SYSTEM
DATA & INPUT
AI MODEL
2. USE PHASE
OUTPUTS

https://oecd.ai/en/ai-principles
2024

Design and training of the system may continue in downstream uses (fine-tuning, continuous learning)

What is an Intelligent System?

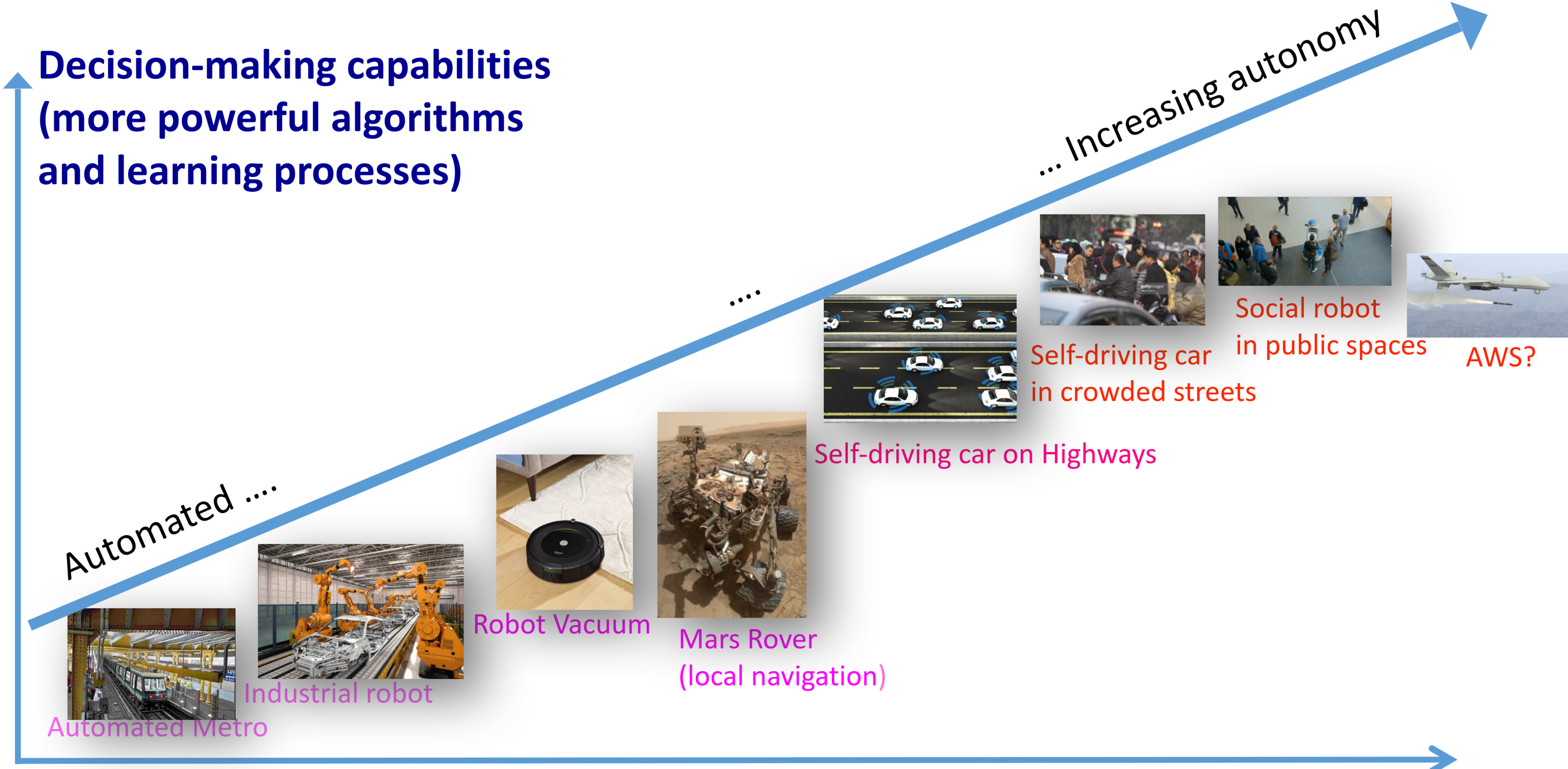
- A computational “intelligent” system is an organized set of algorithms designed by humans, using data to solve complex problems in complex situations.
- The system might use statistical methods for data classification (e.g., deep learning) and improve its performance by evaluating and optimising its decisions (e.g., reinforcement learning).
- Such systems could be regarded as “autonomous” in a given domain and for given tasks, as long as they are capable of making decisions to accomplish their tasks without human intervention, despite variability in operating conditions within this domain.

Machine Autonomy

Task determined and defined by humans

- Autonomy is the capacity of an agent to determine and achieve its actions by its own means
- Autonomy is related to the agent's capability to adapt to environment/task variations
- **Attainable machine autonomy** is relative to task and environment complexity and variability.

- Operational autonomy: Perception, navigation, motion, manipulation, to achieve defined goals (Go to a position, grasp an object, …)
- Decisional autonomy: ability to assess situations and to devise **action plans** for achieving tasks and fulfilling objectives



Decision-making capabilities (more powerful algorithms and learning processes)

Automated ....

....

... Increasing autonomy

Automated Metro

Industrial robot

Robot Vacuum

Mars Rover (local navigation)

Self-driving car on Highways

Self-driving car in crowded streets

Social robot in public spaces

AWS?

Environment and task diversity, complexity, uncertainty

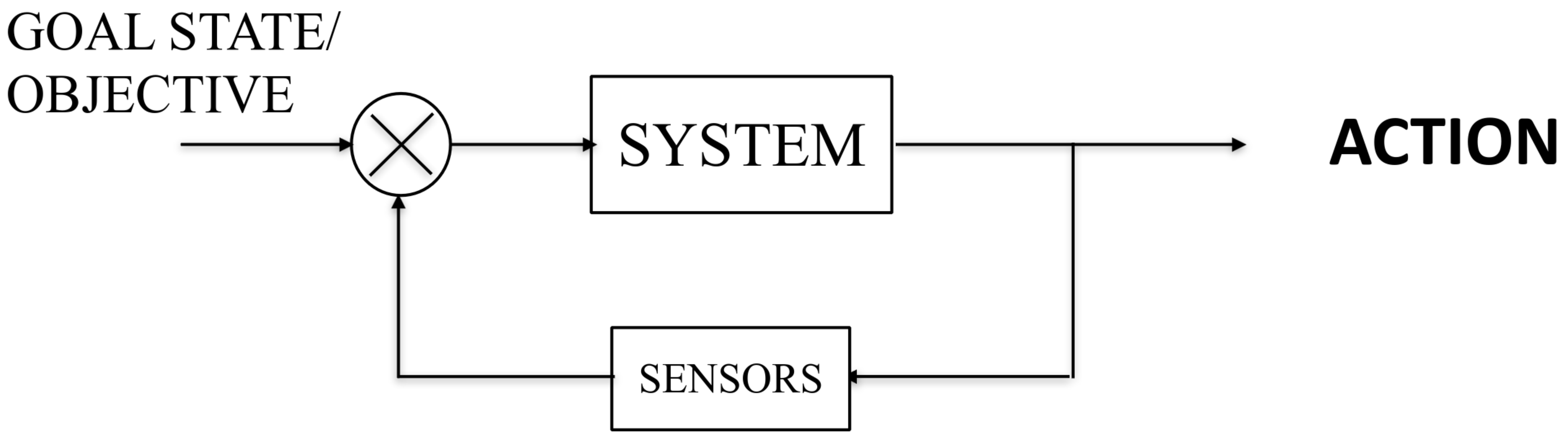
Autonomy is Related to Complexity

- Increasing autonomy is pushing the boundary within which the system can operate with its own capacities.

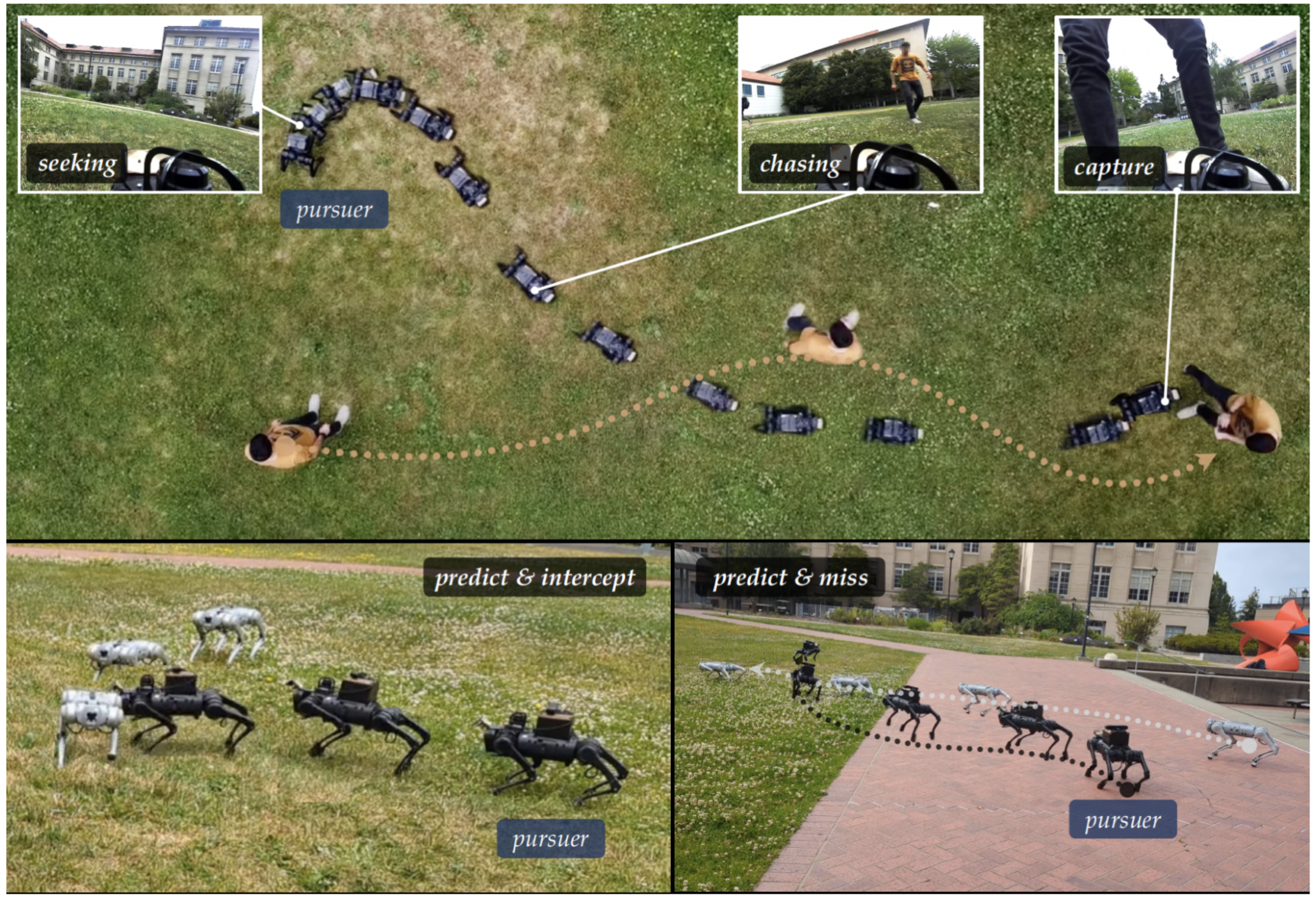
Attainable Autonomy
Complexity of environment
Complexity of task
Attainable Autonomy
Complexity of environment
Complexity of task
Attainable Autonomy
Complexity of environment
Complexity of task

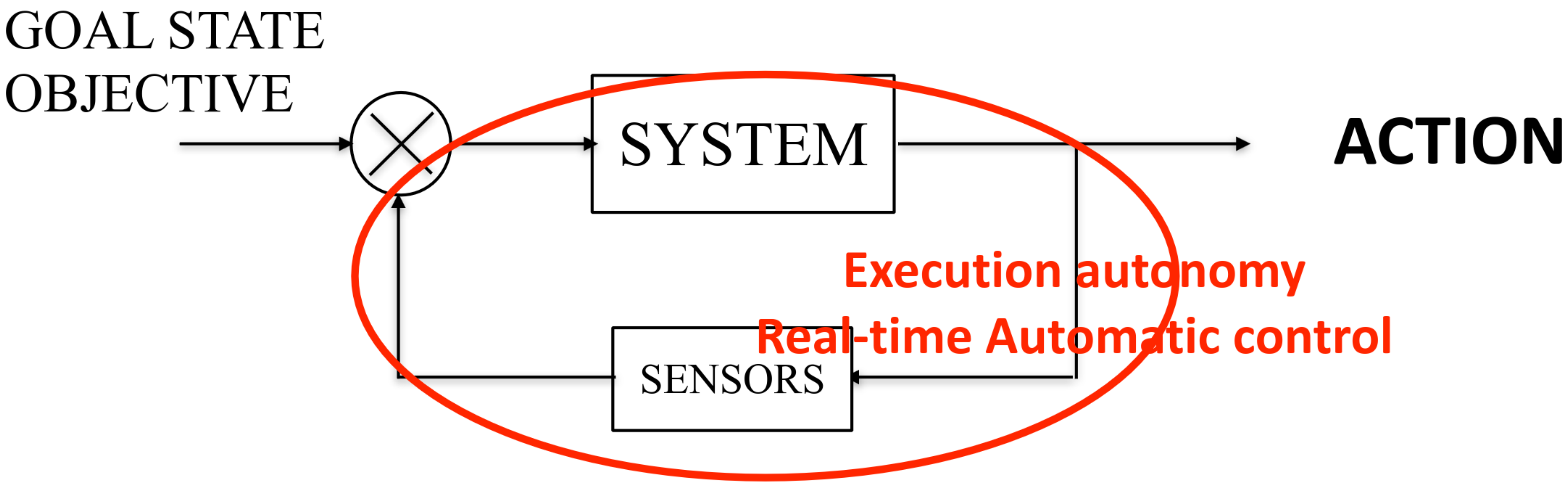
Pushing the boundaries: more advanced programming, or learning

Automated Feedback System



Execution Autonomy





GOAL STATE
OBJECTIVE

SYSTEM

ACTION

SENSORS

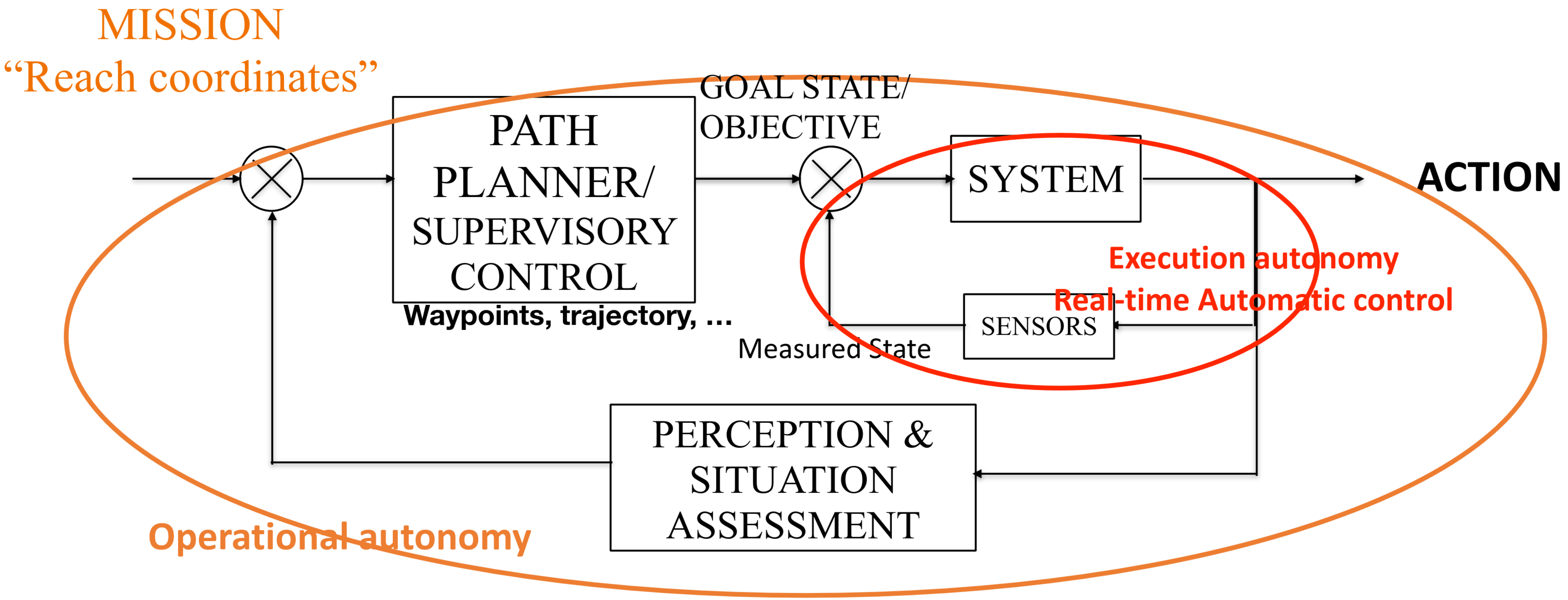
Execution autonomy
Real-time Automatic control

System behavior can be learned (e.g., reinforcement learning

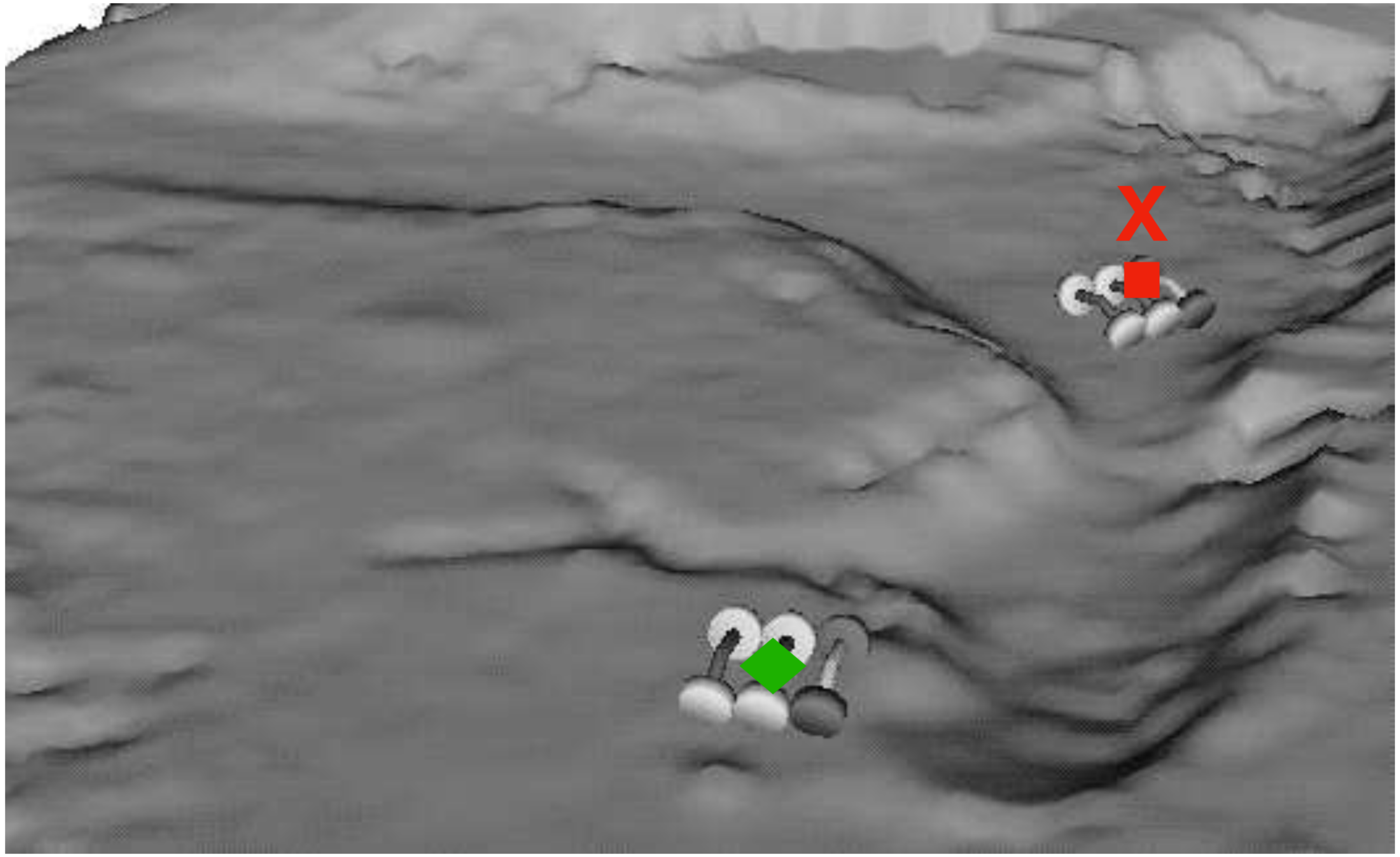
A. Bajczy et al. Learning Vision-based Pursuit-Evasion Robot Policies. arXiv:2308.16185v1 [cs.RO] 30 Aug 2023

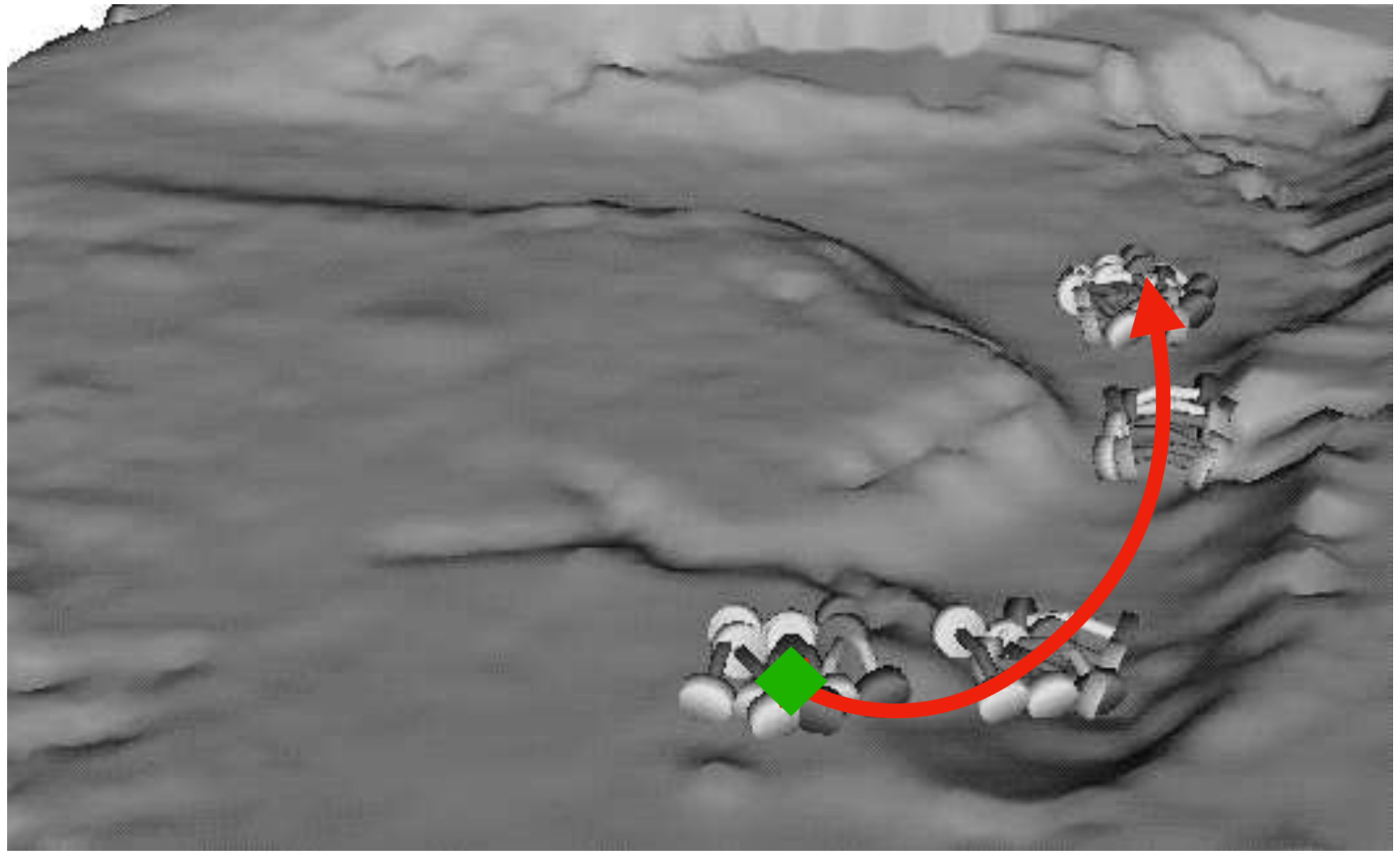
Operational Autonomy

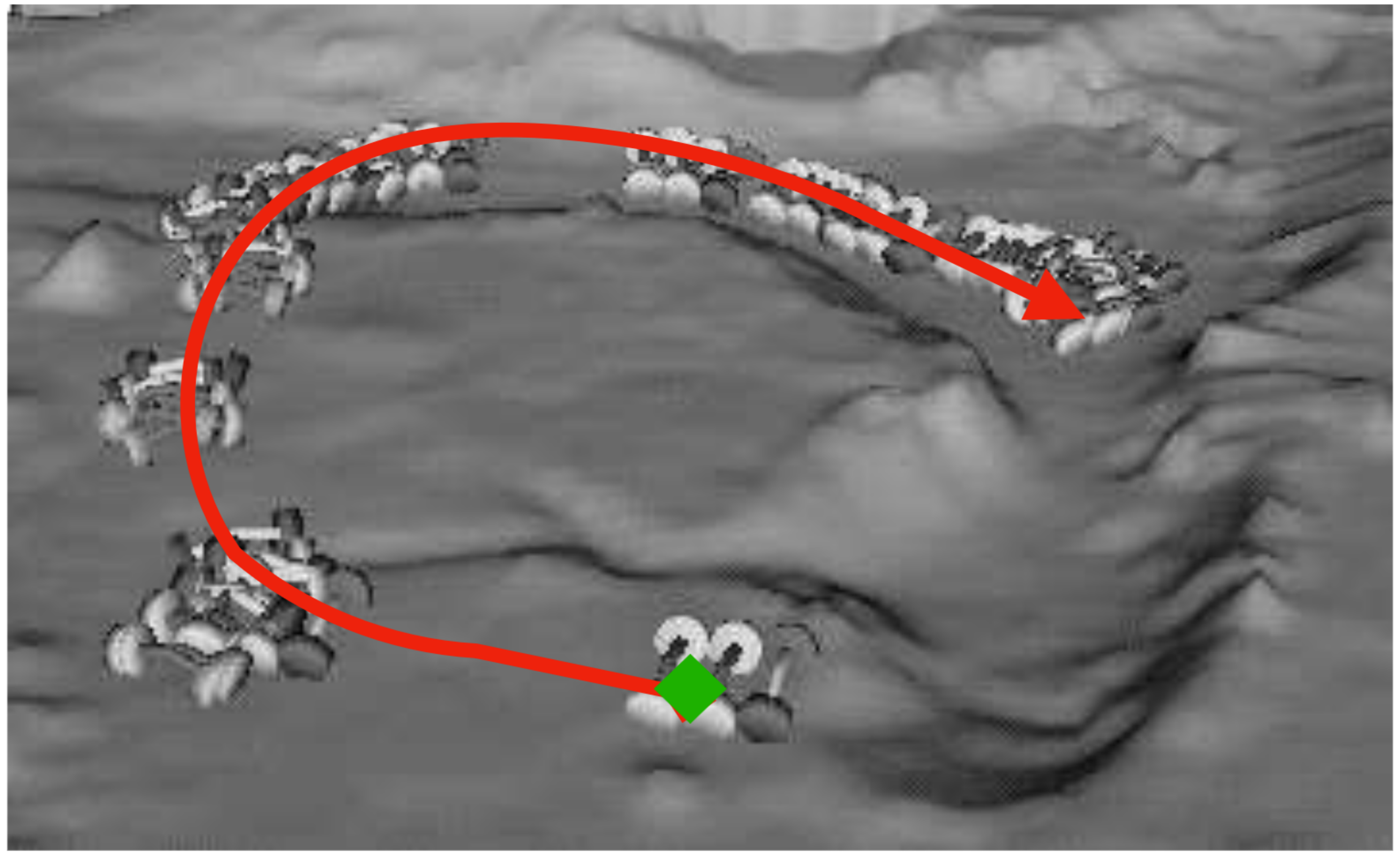
Every function can be based on AI techniques including machine learning



Operational Autonomy





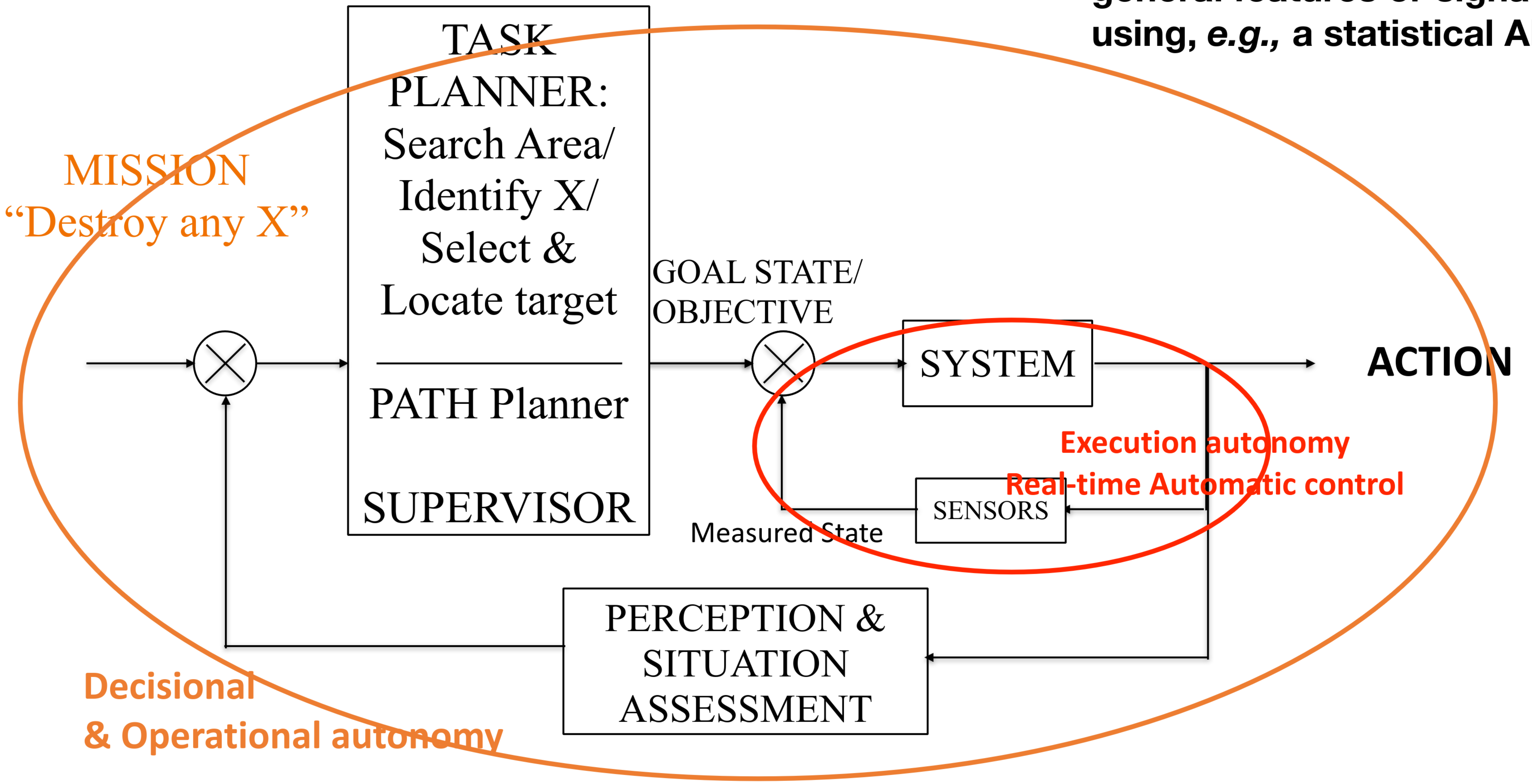


Mission: Goto X

X defined by location coordinates or by specific features

From Operational Autonomy to Decisional Autonomy

X defined and recognised by general features or signature, using, *e.g.*, a statistical AI model



MISSION
"Destroy any X"

TASK PLANNER:
Search Area/
Identify X/
Select &
Locate target

PATH Planner

SUPERVISOR

GOAL STATE/
OBJECTIVE

SYSTEM

ACTION

SENSORS

Measured State

Execution autonomy
Real-time Automatic control

PERCEPTION &
SITUATION
ASSESSMENT

Decisional
& Operational autonomy

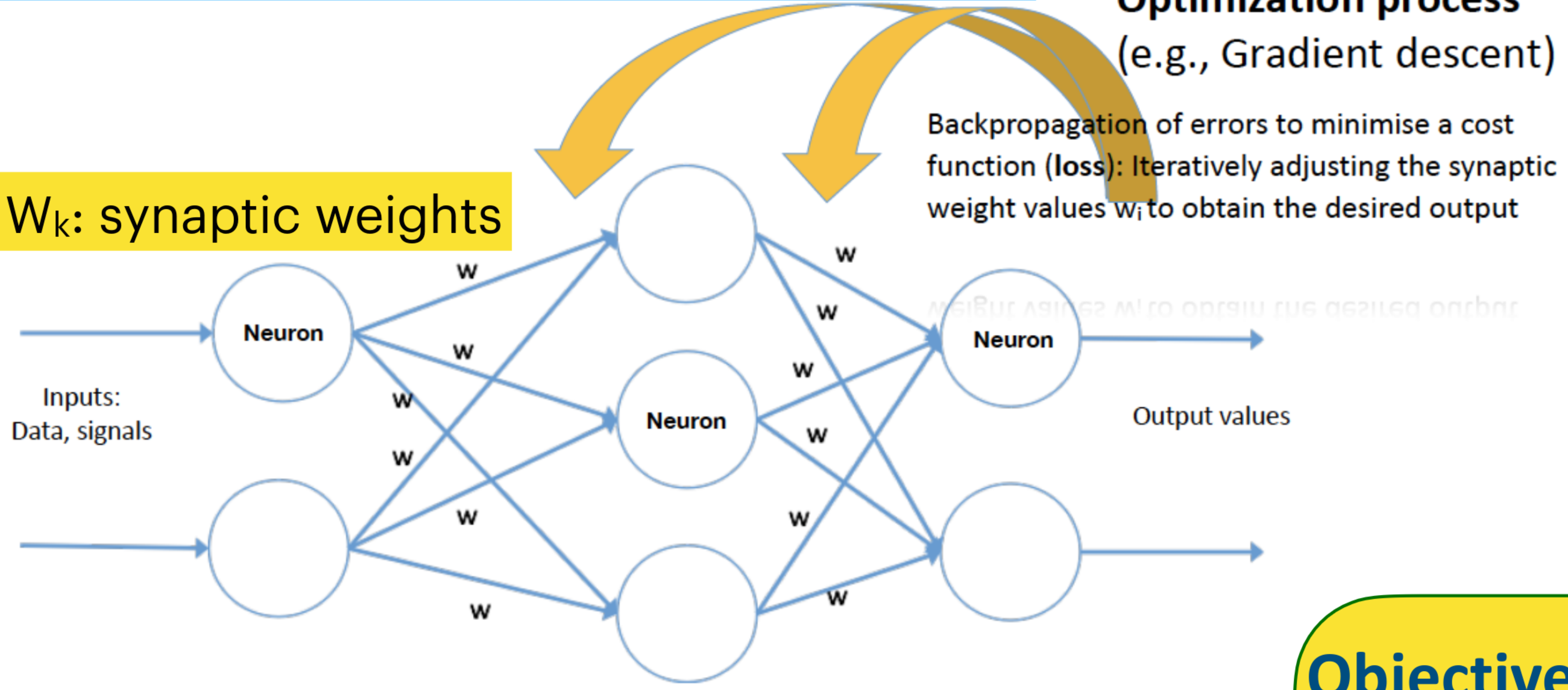
Statistical Machine Learning Methods

Supervised Learning

Optimization process
(e.g., Gradient descent)

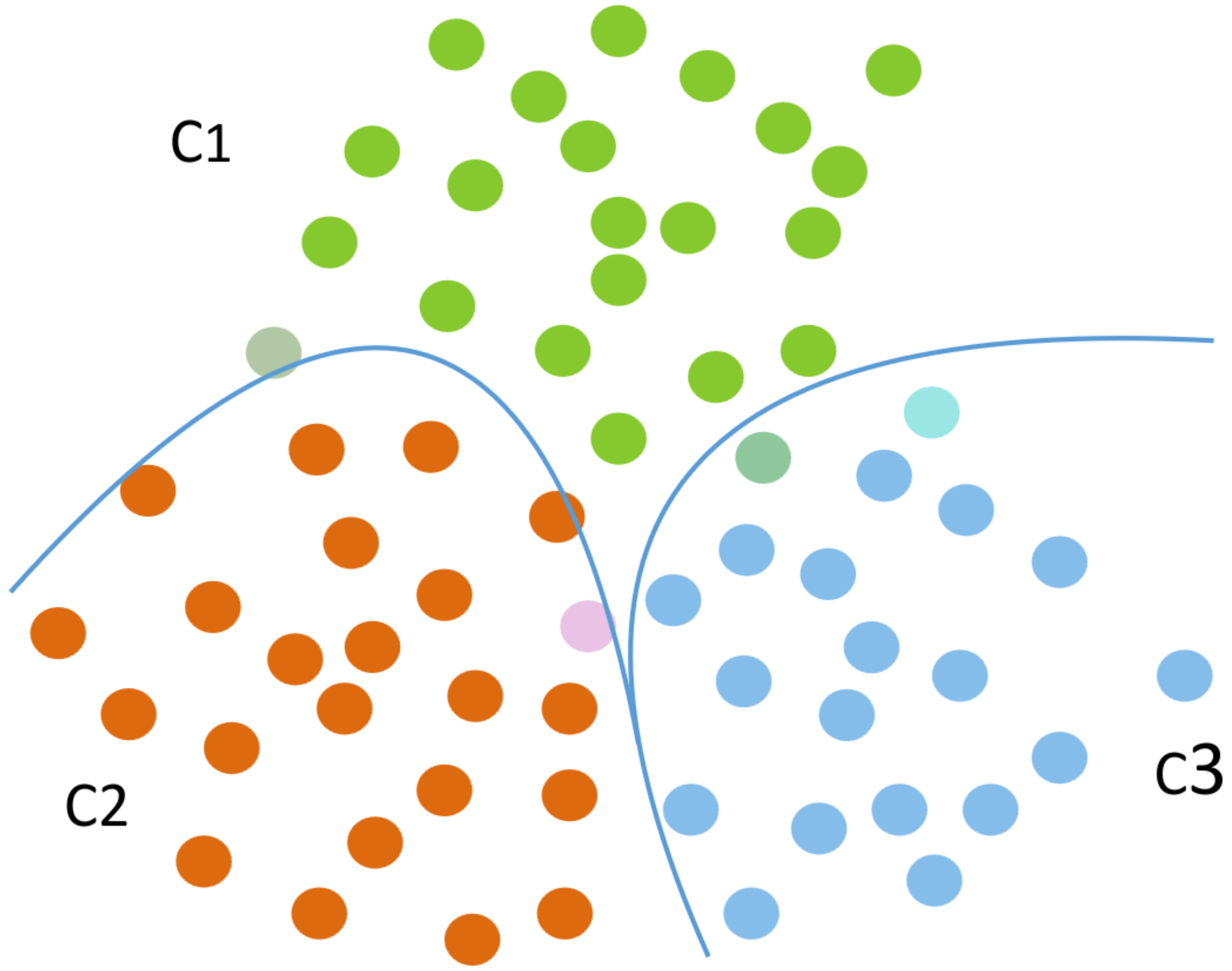
Backpropagation of errors to minimise a cost function (**loss**): Iteratively adjusting the synaptic weight values w_i to obtain the desired output

W_k: synaptic weights



Millions or billions of parameters depending on network size

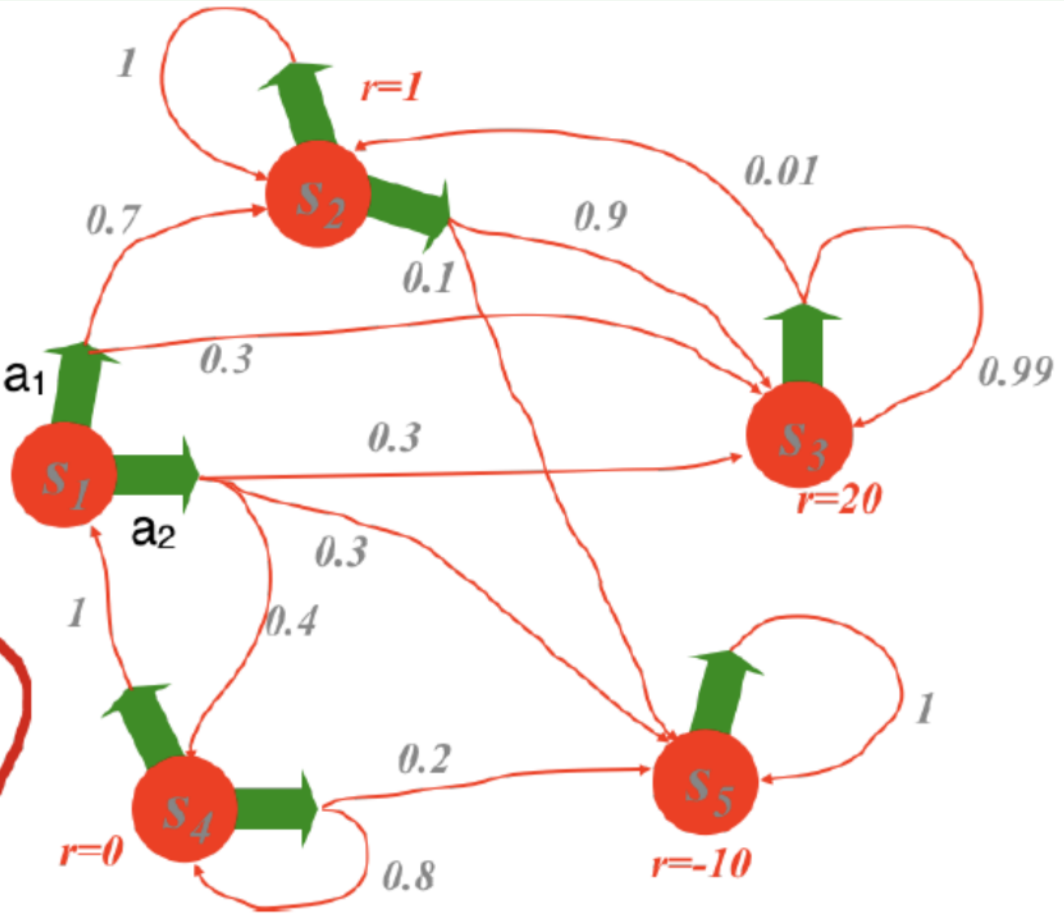
- Finding regularities in data.
- Clustering, classification: K-means, gaussian mixtures, hierarchical clustering, spectral clustering, etc.
- Training data $\{(x_i, y_i)\}/x_i \in \mathbb{R}^p$

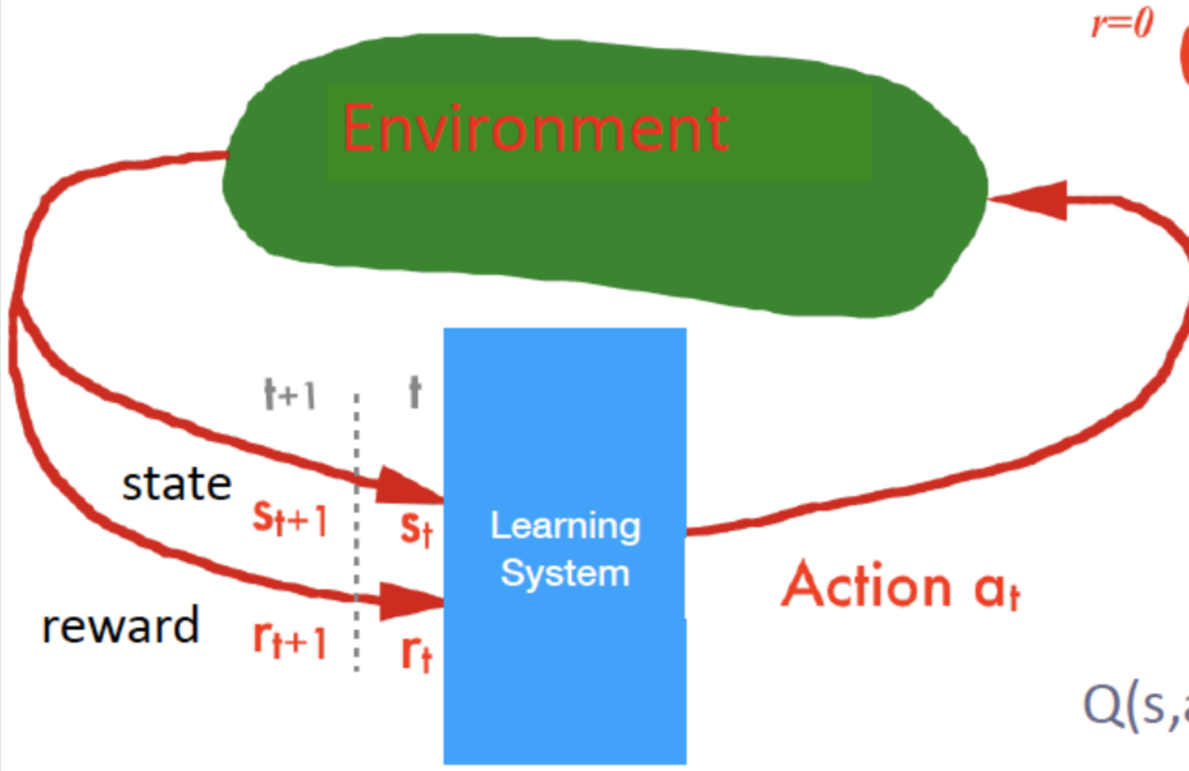


Objective:
Computation of a statistical model of data features to produce categories and predict classes of new inputs

Unsupervised Learning

Reinforcement Learning
Maximizing total expected reward

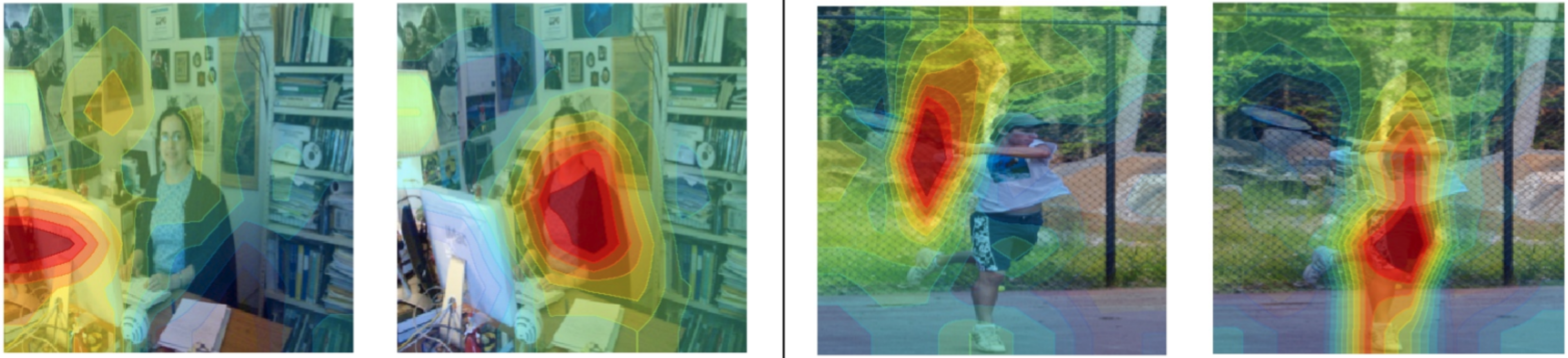




$$Q(s,a) = R(s) + \gamma\sum_{s'}P(s,a,s')\max_{a'}Q(s',a')$$

Reinforcement Learning

Machine Learning Limitations
Data Bias



Wrong

Right for the Right Reasons

Right for the Wrong Reasons

Right for the Right Reasons

Baseline:
*A **man** sitting at a desk with a laptop computer.*

Our Model:
*A **woman** sitting in front of a laptop computer.*

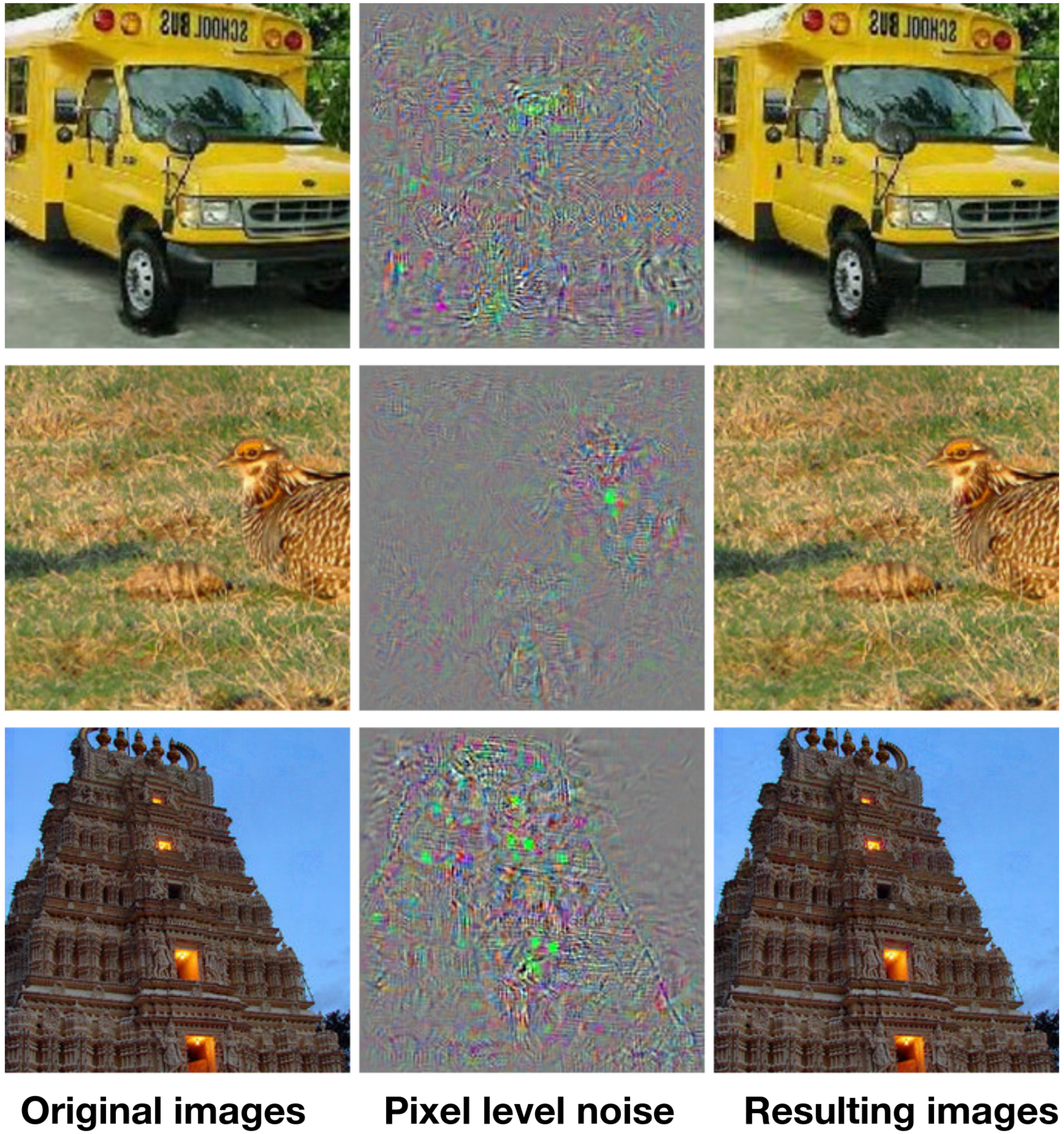
Baseline:
*A **man** holding a tennis racquet on a tennis court.*

Our Model:
*A **man** holding a tennis racquet on a tennis court.*

Women also Snowboard: Overcoming Bias in Captioning Models.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, Anna Rohrbach. ECCV 2018

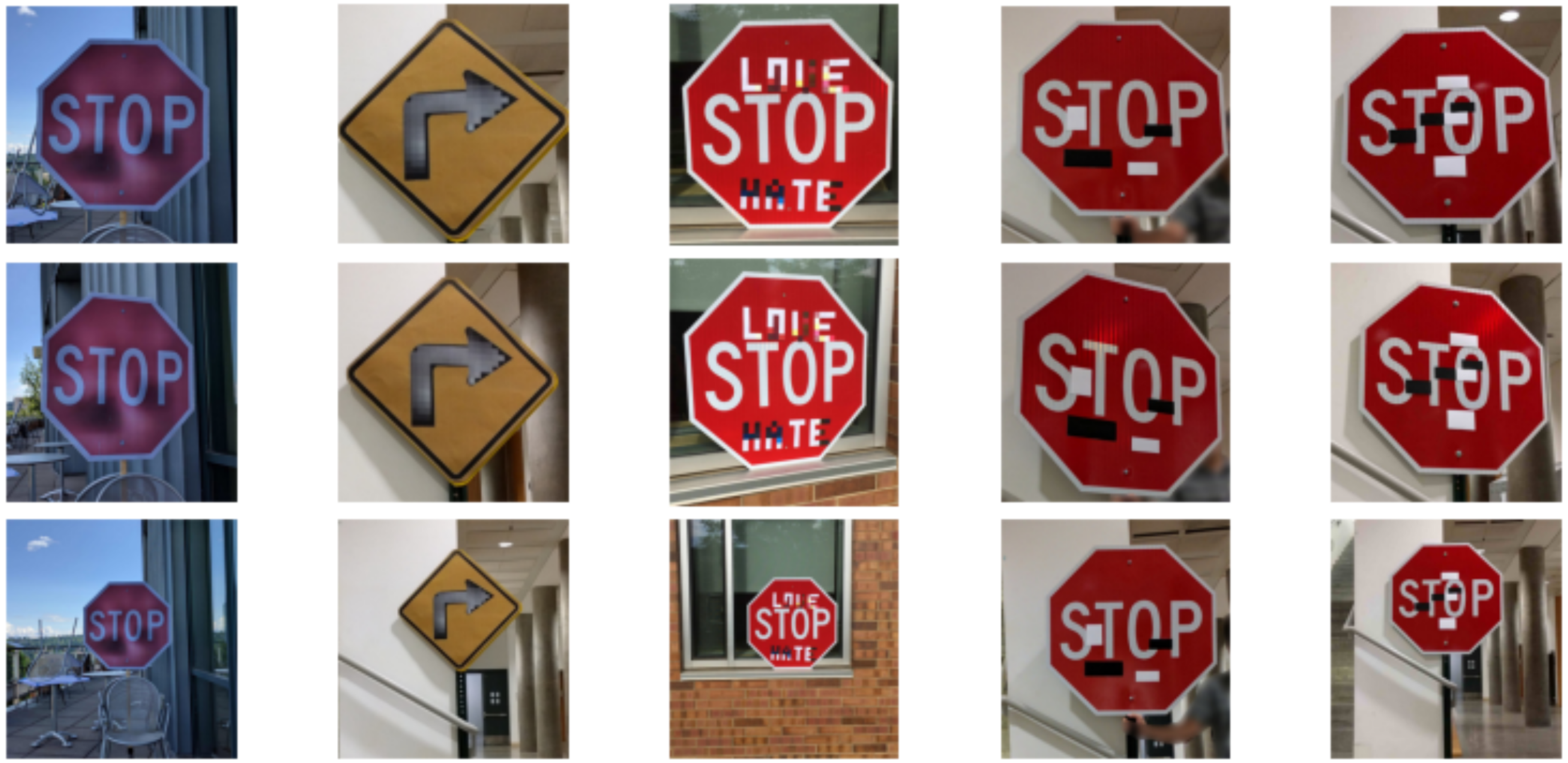
Deep Learning Limitations: Sensivity to noise



Intriguing properties of neural networks
Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus
https://doi.org/10.48550/arXiv.1312.6199

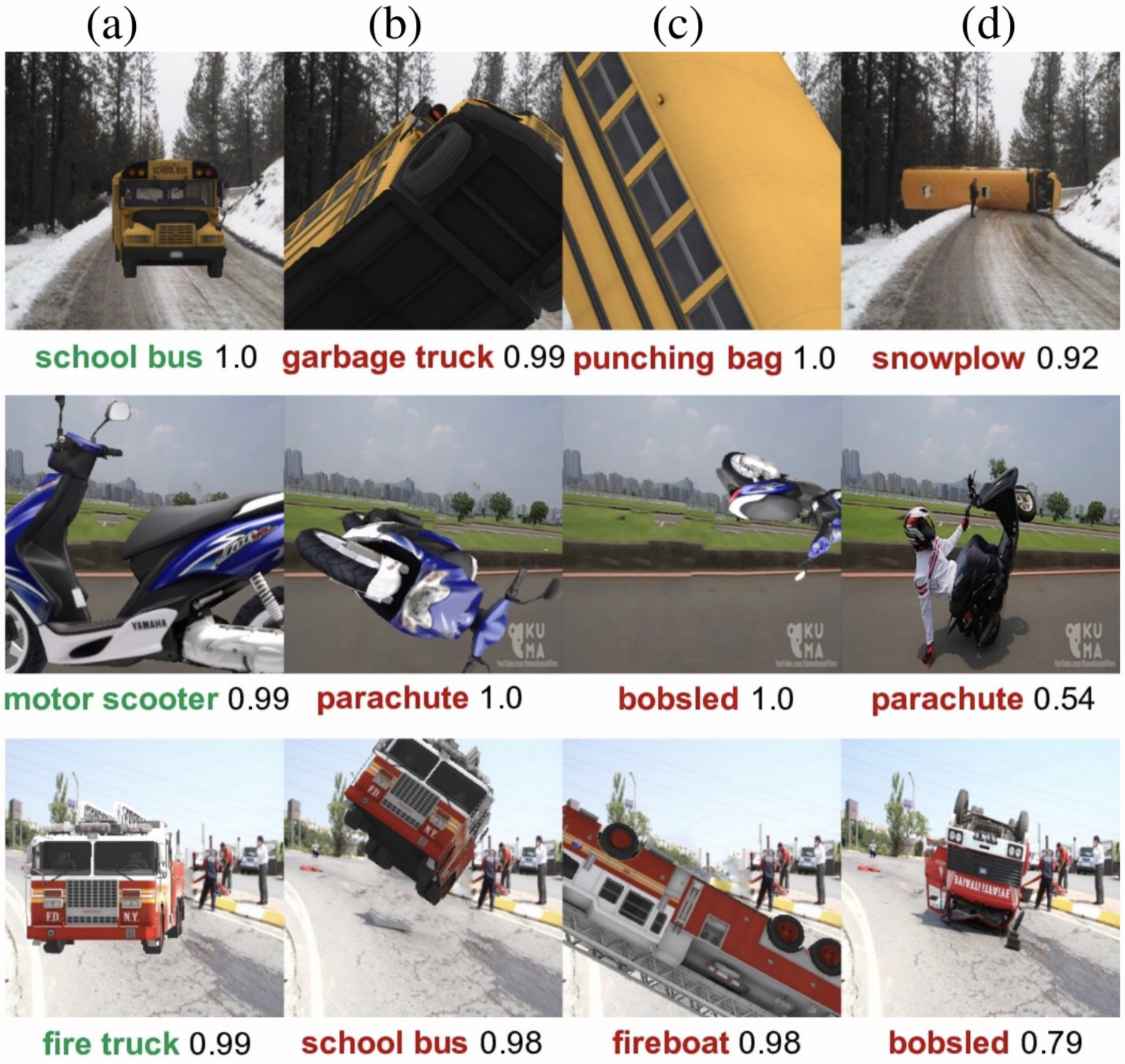
Images in the right column are predicted to be an “ostrich”

Deep Learning Limitations
Lack of Robustness



Targeted physical perturbation experiment
The misclassification target was Speed Limit 45.

Robust Physical-World Attacks on Deep Learning Models K. Eykholt et al. CVPR 2018.



Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects. Michael A. Alcorn et al., CVPR 2019

Summary Issues with Statistical Machine Learning

- Black box: millions/billions of parameters
- Data bias: quality and representativeness of data
- Design Bias: Hyperparameters, architecture choices, optimisation algorithms
- Spurious correlations, confabulations, mixture of false and true information
- Unpredictability and sensitivity to inputs
- Absence of causality between data and results
- No or little explicability
- No semantics or grounding in the real world, no abstraction
- Misalignment
- No solid verification and validation processes or qualification of results
- Environmental cost

Autonomous Weapons : Technically Framing Autonomy?

- Robustness: certifying perception and decision systems
- No continual learning
- Limiting system’s degrees of freedom for interpretation (more precise target characterization)
- Limiting the global space-time mission domain (reducing situation evolution)
- Limiting mission dynamics uncertainty (difference between situation at activation and situation during execution) through global monitoring (loitering ammunitions, …)
- Guaranteeing human control after deployment (needs situation observation and mission abort capacities)
- Explainability, Forensics and audibility for accountability

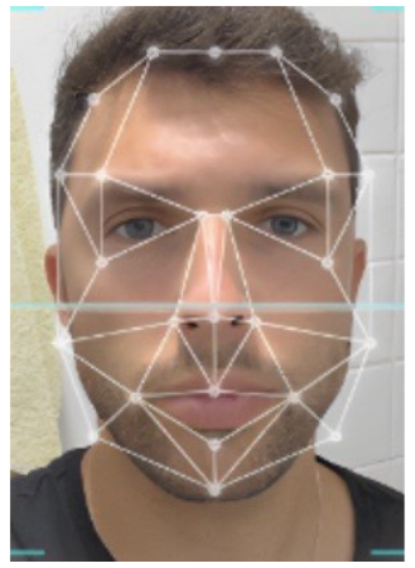


Image credits:
The Washington Post
Medium,
Josaphat Musamba
Lankaweb
Military Aerospace

Do LAWS raise ethical issues?

Machine Autonomy and Ethics

- Machines operate at the **computational level**: global contextual knowledge and **semantic situation understanding** is beyond machines capacities. Machines don’t understand what human beings are, what human dignity means.
  - Is it ethical/acceptable that a machine makes a decision of life and death over humans?

- Machine decisions and behavior are the result of (imperfect) algorithms and computational processes
  - Risk of unpredictable harm and behavior

- Machines don’t have moral agency. They cannot make ethical decisions based on moral judgement. They can only apply **decision criteria** provided through programming
  - Pre-established criteria cannot provide for contextual decisions. No discernment, no temperance.

Requirements for Trustworthy AI

High-Level Expert Group on AI (EU) - April 2019

European Commission

Independent
High-Level Expert Group on
Artificial Intelligence
set up by the European Commission

AI

Ethics Guidelines
for Trustworthy AI

1. **Human agency and oversight -** human control
2. **Technical robustness and safety** - general safety, accuracy, fall back plan, reliability and reproducibility, resilience to attack and security
3. **Privacy and data governance** - quality and integrity of data, and access to data
4. **Transparency** - Including traceability, explainability and communication
5. **Diversity, non-discrimination and fairness** - avoidance of unfair bias, accessibility and universal design, stakeholder participation
6. **Societal and environmental wellbeing** - Including sustainability and environmental friendliness, social impact, society and democracy
7. **Accountability** - auditability, minimisation and reporting of negative impact, trade-offs and redress.

Tool: Assessment List for Trustworthy AI - ALTAI

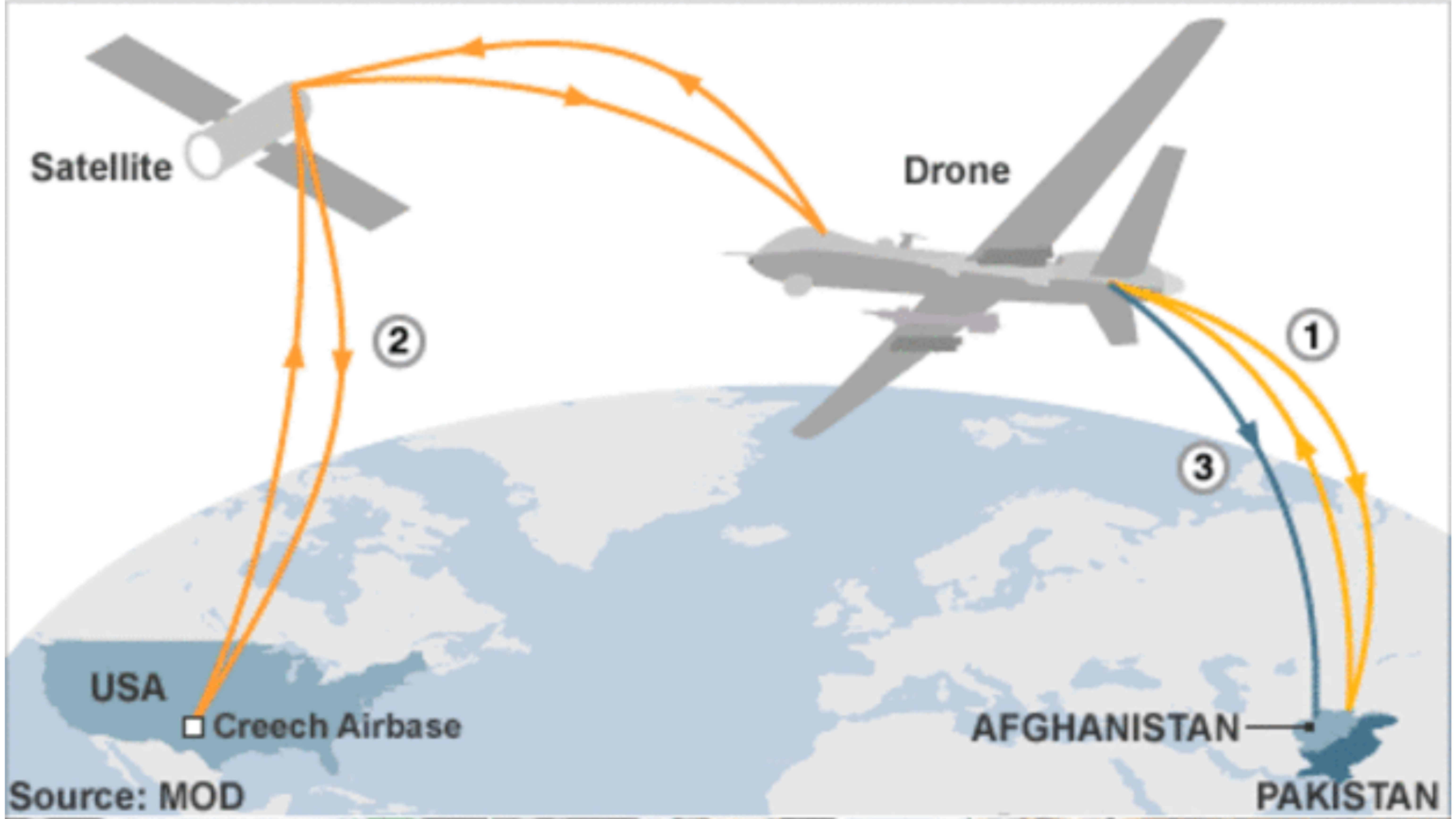
https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

Meaningful Human Control

Preserving **meaningfu**l human control over the use of (lethal) force, that is: humans not computers and their algorithms should ultimately remain in control of, and thus morally responsible for relevant decisions about (lethal) military operations. (Proposed by the NGO “Article 36”, 2015)

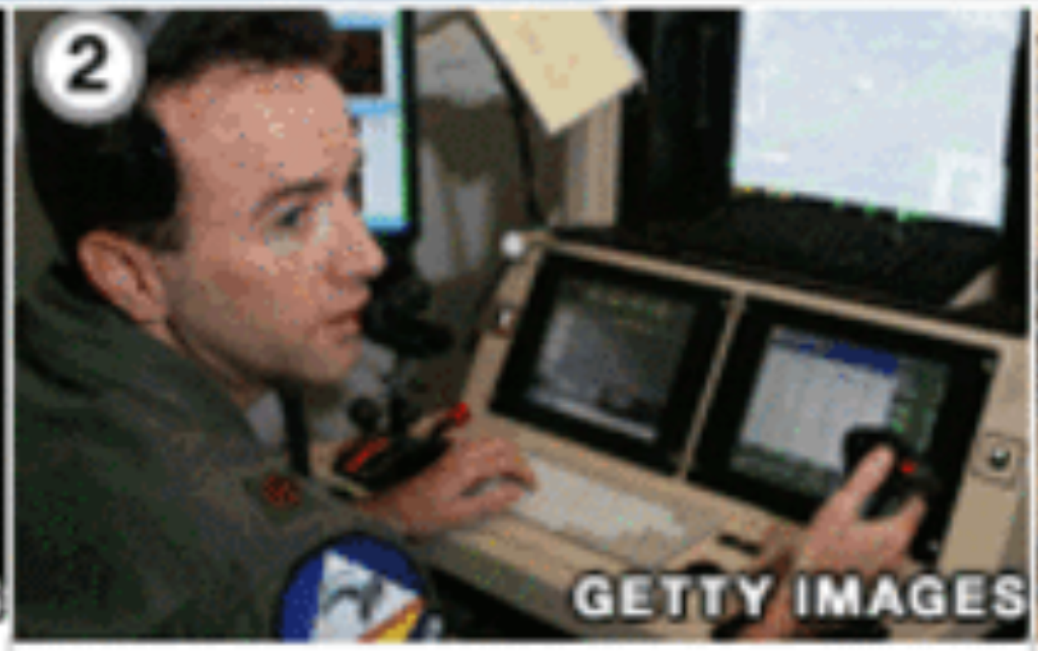
- Definition of meaningful?
- Does meaningful human control contradict the concept of AWS (permanent communications, permanent monitoring)?

How drones work



Source: MOD

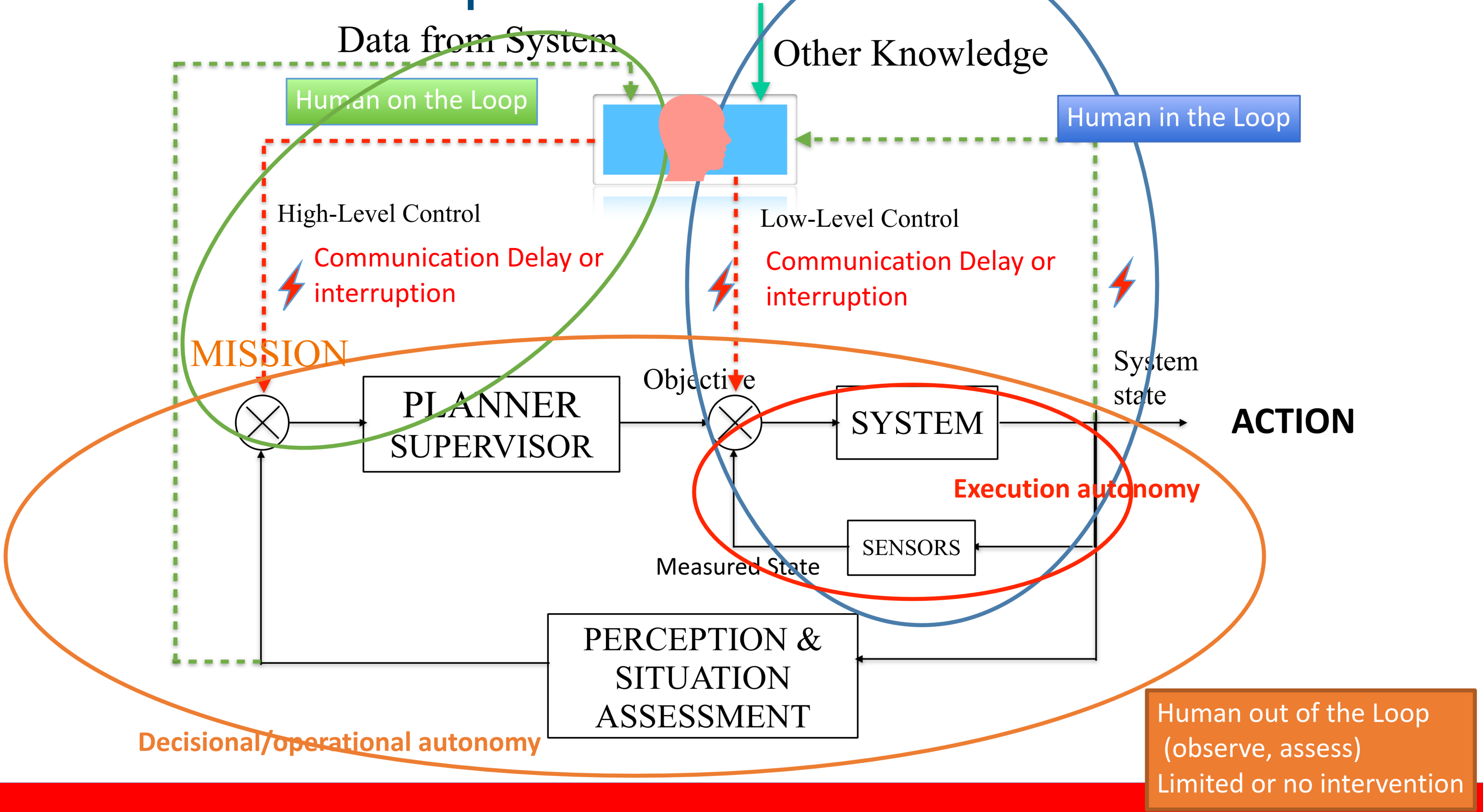
Drone take-off and landing controlled locally



Drone flown remotely from US airbase

Images relayed to troops on the ground

Humans and Loops



Human out of the Loop
(observe, assess)
Limited or no intervention

Shared Authority Between Human & Machine

- Machine control
  - Limited decision-making capacities
  - Sensing uncertainties
  - Limited situation assessment
  - Rational decisions
  - Reactivity
  - No morality, no understanding of values

- Human control
  - Limited attention span
  - Limited perception field
  - Stress and emotions
  - Global situation awareness
  - Moral judgment

Interaction Problems

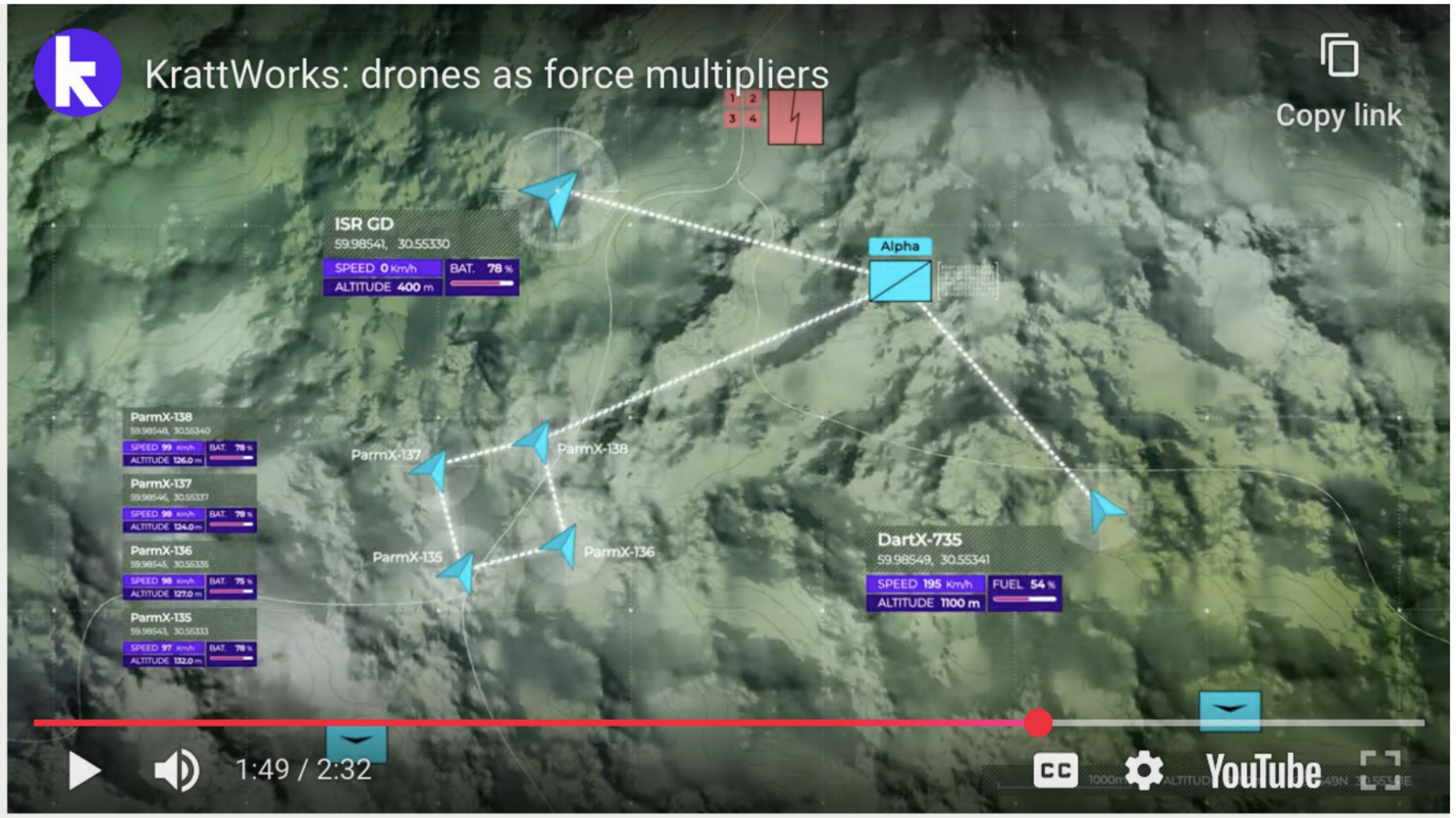
Automation bias: overconfidence in the machine

Surprises: ignorance of exact state in case of take-over

Moral buffer: machine responsibility vs. human responsibility

Decision Support Systems

- Decision Support Systems (DSS) are tools and systems for assisting human decision-making.
- Situation assessment, visualisation, scenarios, simulation, prediction, target identification, prioritisation, recommendations for action, …
- Sophisticated interfaces
- Same limitations as AI systems (on which they are based)
- Might lead to misinformed human decisions and reduce huma role to validation



A promotional video from Krattworks depicts scenarios in which the company's drones augment soldiers on offensive maneuvers. KRATTWORKS

Are LAWS compatible with IHL?

International Humanitarian Law

- Foundation: *limiting the use of violence in armed conflicts by sparing those who do not or no longer directly participate in hostilities*
- Principles:
  - the principle of **humanity** (the "elementary" considerations of humanity being reflected and expressed in the Martens clause)
  - the principle of **distinction** between civilians and combatants, and between civilian objects and military objectives
  - the principle of **proportionality** prohibiting attacks "which 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"
  - the principle of **military necessity** (from which flows the prohibition of superfluous injury and unnecessary suffering).

Source ICRC: https://casebook.icrc.org/a_to_z/glossary/fundamental-principles-ihl

AWS and IHL?

- Lack of semantics and understanding
- Lack of globally contextual decision making
- Unpredictability (situation assessment, swarm behavior)
- No moral agency

- **Humanity?**
- **Distinction?**
- **Proportionality?**
- **Necessity?**

- Responsibility and accountability ?

- Easy access to technology and dissemination

Towards a Regulation?

- November 2024: 161 member states of the UNGA voted on a resolution that raises concern about the “*negative consequences and impact of autonomous weapon systems on global security and regional and international stability, including the risk of an emerging arms race, of exacerbating existing conflicts and humanitarian crises, miscalculations, lowering the threshold for and escalation of conflicts and proliferation, including to unauthorised recipients and non-State actors.*”
- 2025: 129 states support call to negotiate a treaty that prohibits and regulates autonomous weapons systems.
- The GGE should conclude with a binding instrument proposal in 2026
- Some actors consider moving the issue to the UNGA in case of failure