In Search of a Foundation for Next Generation Autonomous Systems (Can We Trust Autonomous Systems?)

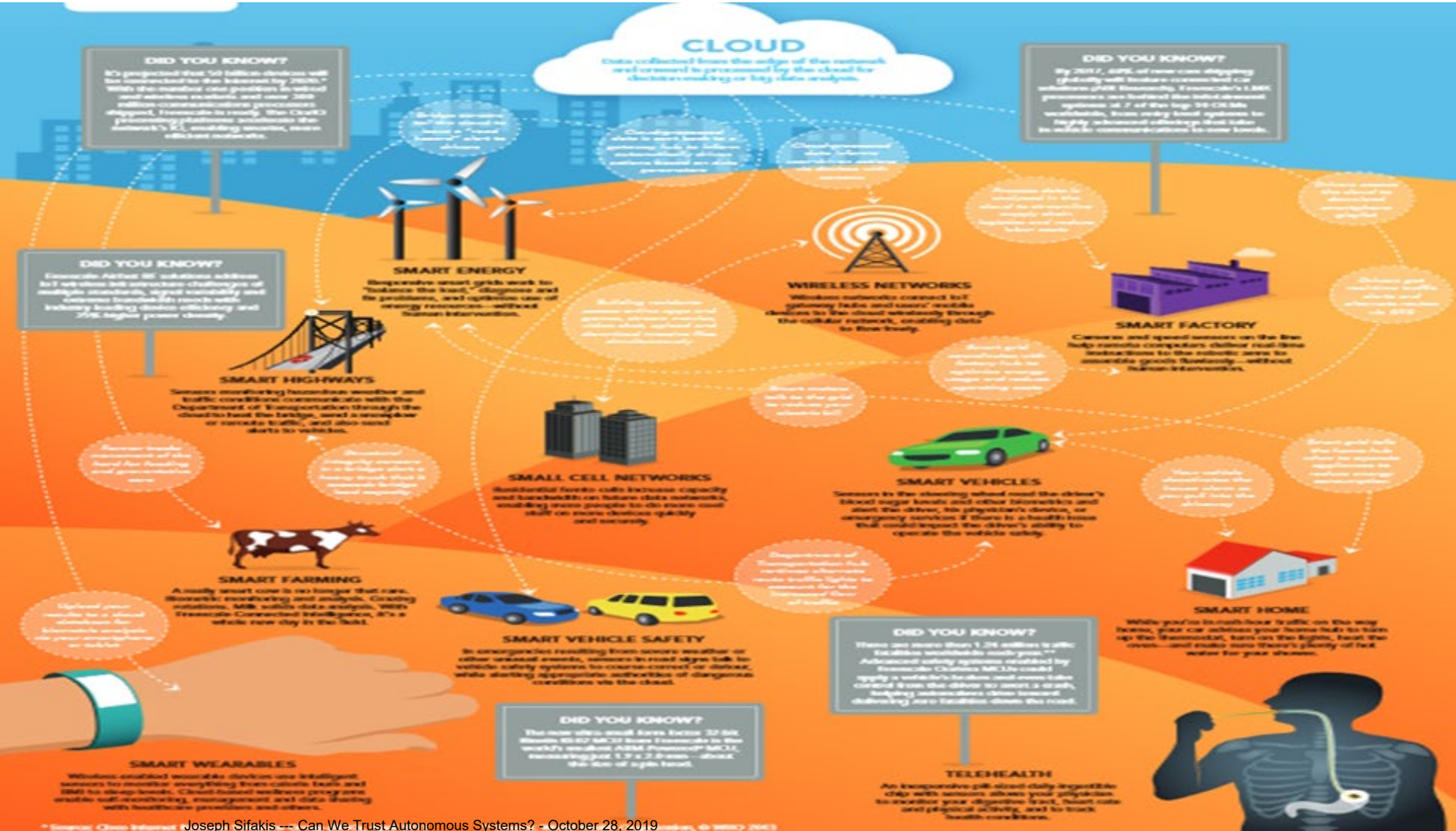
ATVA 2019

Taipei, October 28, 2019

Joseph Sifakis
Verimag Laboratory

Next-generation autonomous systems – The IoT Vision

The IoT allows objects to be sensed or controlled remotely across a network infrastructure, achieving more direct integration of the physical world into computer-based systems, and resulting in improved efficiency and predictability.



Next-generation autonomous systems – The IoT Vision

The Internet of Things

Industrial IoT

Autonomous

Rules can be changed, but human-driven changes are external to normal behavior

Autonomous transport systems
Industry 4.0
Smart grids

Human IoT

Interactive

People's explicit or arbitrary actions dynamically trigger control sequences or rule changes

Intelligent services
Semantic web

Next-generation autonomous systems – Main Characteristics

Next-generation autonomous systems emerge from the needs to further automate existing organizations by progressive and incremental replacement of human operators by autonomous agents.

- Such systems are often critical and should exhibit "broad intelligence" by handling knowledge in order to
  - Manage dynamically changing sets of possibly conflicting goals – this reflects the trend of transitioning from "narrow" or "weak" AI to "strong" or "general" AI.
  - Cope with uncertainty of complex, unpredictable cyber physical environments.
  - Harmoniously collaborate with human agents e.g. "symbiotic" autonomy.

- Serious limitations to meeting criticality requirements e.g.
  - No trustworthiness assurance techniques for learning-enabled components;
  - Poor trustworthiness of the network infrastructure required to deal with geographic distribution and mobility e.g. security issues, impossibility to guarantee response times
  - Overwhelming complexity due to highly dynamic behavior e.g cyber physical agents, and unpredictability

Autonomous Vehicles are an emblematic topical case raising cutting-edge challenges and involving huge economic stakes and deep societal impact.

Next-generation autonomous systems – New Trends

- In contrast to the aerospace and rail industries,
  - AV manufacturers have not followed a "safety by design" concept; they adopt "black-box" ML-enabled end-to-end design approaches.
  - AV manufacturers consider that statistical trustworthiness evidence is enough - "*I've driven a hundred million miles without accident. Okay, that means it's safe.*"
  - Public authorities allow "self-certification" for autonomous vehicles.
  - Critical software can be customized by updates – Tesla cars software may be updated on a monthly basis. (*)

(*) Aircraft are certified as products – SW or HW components cannot be modified!

- Prevailing attitudes about the lack of rigorous design methods
  - Blunt realism. Charge ahead, accept the risks: the benefits will be so great!
  - Blind faith in empirical methods. Rigorous approaches are inherently inadequate; complex problems can be solved only by empirical methods.
  - Unbridled optimism. We have the right tools, it's just a matter of time.(*)

(*)"I almost view [autonomous cars] as a solved problem. We know what to do, and we'll be there in a few years." E. Musk, Nvidia Technology Conference, March 2015.

Next-generation autonomous systems – The Issue of Trust

	Systems Engineering is facing a huge gap, moving				
FROM	Small size	Centralized	Automated	Predictable Envt	Elicitable Specs
TO	Complex	Decentralized	Autonomous	Unpredictable Envt	Non-elicitable Specs

We need a new scientific and engineering foundation that cannot be obtained by simply combining existing results developed for more than two decades and focusing mainly on SW systems e.g. Autonomic computing, Adaptive systems, Autonomous Agent Systems and bringsanswers to the following problems:

1. Understand the spectrum of possibilities between Automation and Autonomy
   - What are the technical solutions for enhancing a system's autonomy? - for each enhancement, what are the implied technical difficulties and risks?
   - Principled decision whether we can trust a system to perform a given task.

2. Relate system trustworthiness to knowledge truthfulness about the developed system.

3. Move from traditional system design to "hybrid" design seeking trade offs between trustworthiness of model-based and performance of data-based approaches.

- Autonomous Systems
  - The concept of autonomy
  - The Automation Frontier
- Knowledge Truthfulness
- Design for Trustworthiness and Performance
  - Complexity Issues
  - "Hybrid" design flows
  - Validation
- Discussion

OVERVIEW

The Concept of Autonomy – Find the Differences

Thermostat

Automatic train shuttle

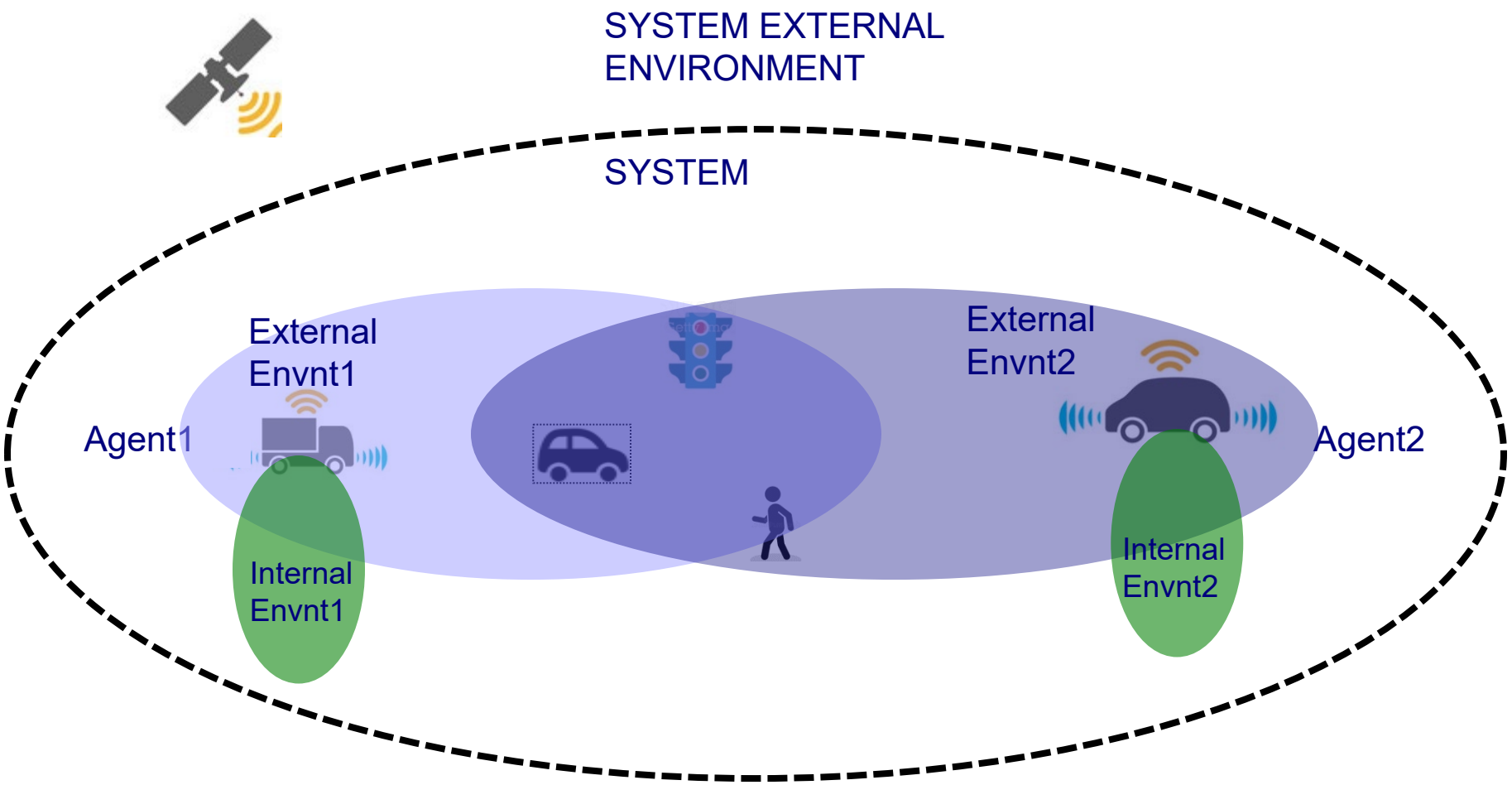
Chess-playing robot

Soccer-playing robot

Robocar

Each system consists of agents acting as controllers on their environment and pursuing individual goals so that the collective behavior meets the system global goals.

The Concept of Autonomy – Basic Definitions



SYSTEM= Agents + Objets + System_Environment

Agents=Agent1+Agent2

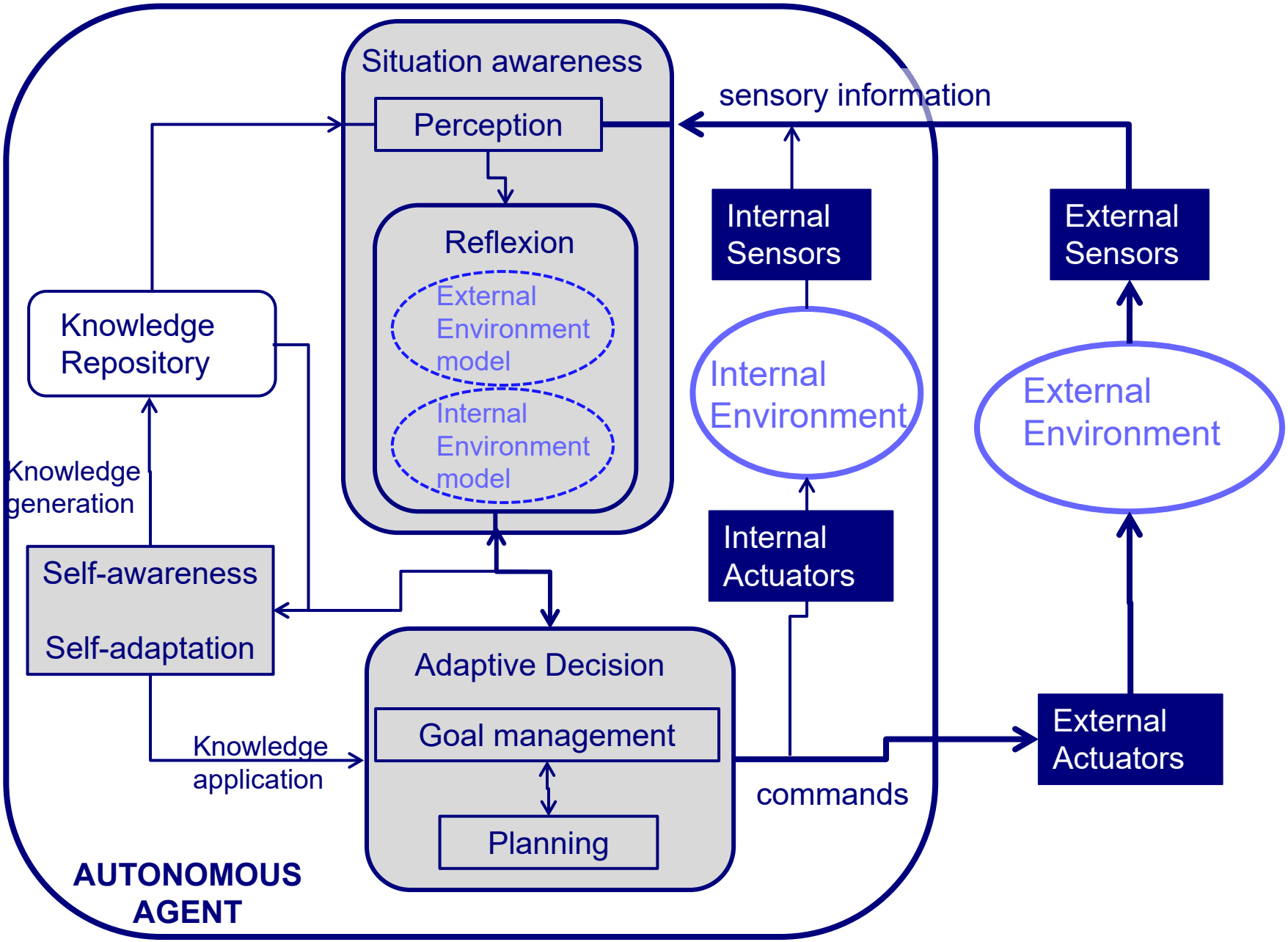
Objects= Traffic_light+Pedestrian+ Human_Driven_car

System_Environment = (External_Envnt1+External_Envnt2)x(Internal_Envnt1+Internal_Envnt2)

The Concept of Autonomy – From Automation to Autonomy

	Environment	Stimuli	Meeting Goals
Thermostat	Room + Heating/cooling device	Temperature	Explicit controller Single goal
Shuttle	Cars + Passengers+ equipment	Static configuration of cars+ State of equipment	Explicit controller + on line adaptation Many fixed goals
Chess robot	Chess board + pawns	Static configuration of pawns	On-line planning+ stored knowledge Dyn. Changing goals
Soccer robot	Regions in the field + Players + Ball	Dynamic configuration of players/ball	On-line planning+ stored/generated knowledge Dyn. changing goals
Robocar	Vehicles/obstacles + Road/communication equipment	Dynamic configuration of vehicles/obstacles + State of equipment	On-line planning+ stored/generated knowledge Dyn. changing goals

The Concept of Autonomy – Architectural Characterization



The Concept of Autonomy – Architectural Characterization

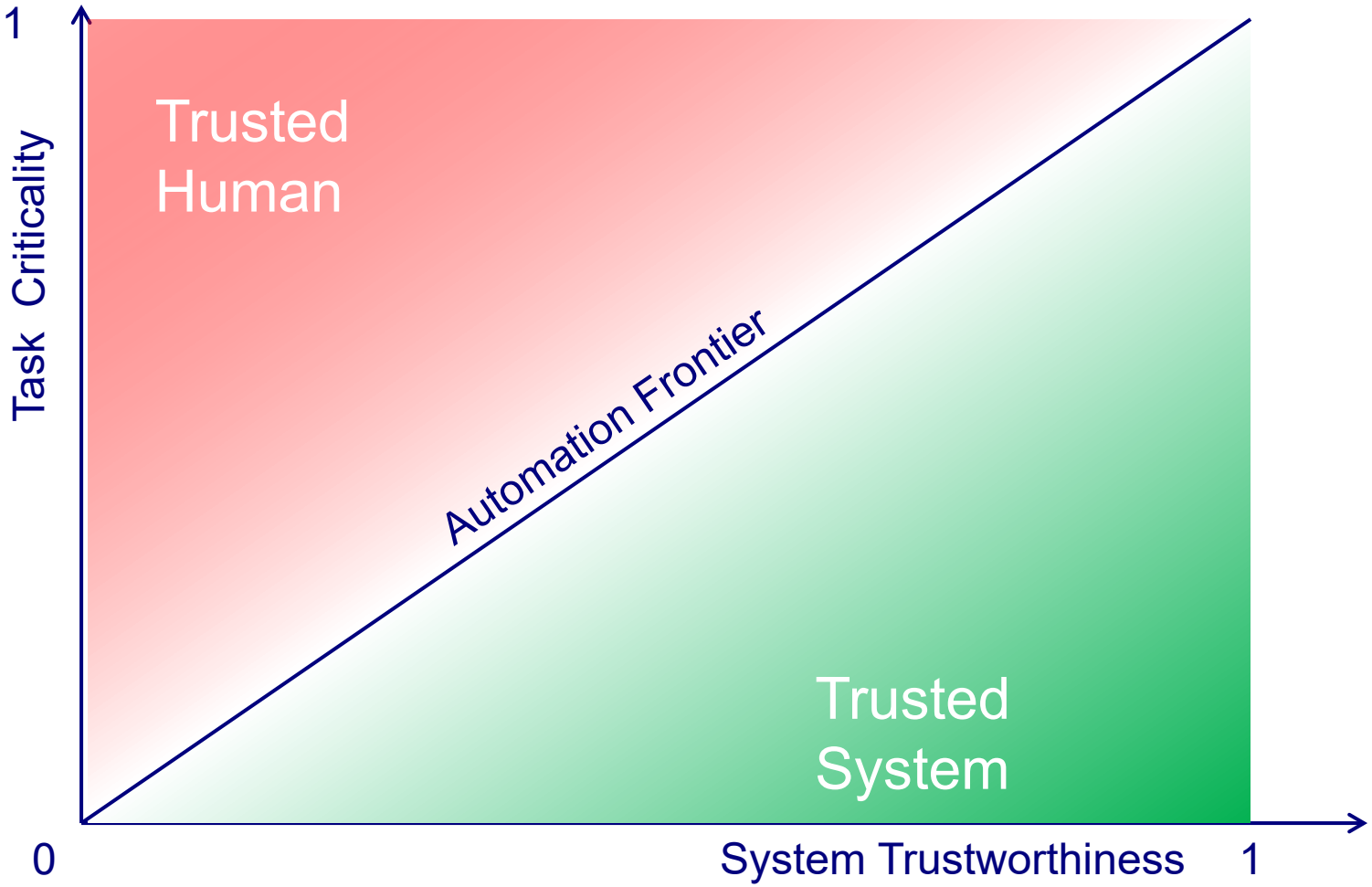
- Autonomy is the capacity of an agent to achieve a set of coordinated goals by its own means (without human intervention) adapting to environment variations. It combines five complementary functions:
  - Perception e.g. interpretation of stimuli, removing ambiguity from complex input data and determining relevant information;
  - Reflection e.g. building/updating a faithful environment run-time model from which strategies meeting the goals can be computed;
  - Goal management e.g. choosing among possible goals the most appropriate ones for a given configuration of the environment model;
  - Planning to achieve a particular goal;
  - Self-awareness/adaptation e.g. the ability to create new situational knowledge and new goals through learning and reasoning.

- These functions are implementation-agnostic.
- Insights on
  - Automation vs. Autonomy;
  - Human-assisted vs. Machine Empowered autonomy.

- Autonomous Systems
  - The concept of autonomy
  - The Automation Frontier
- Knowledge Truthfulness
- Design for Trustworthiness and Performance
  - Complexity Issues
  - "Hybrid" design flows
  - Validation
- Discussion

OVERVIEW

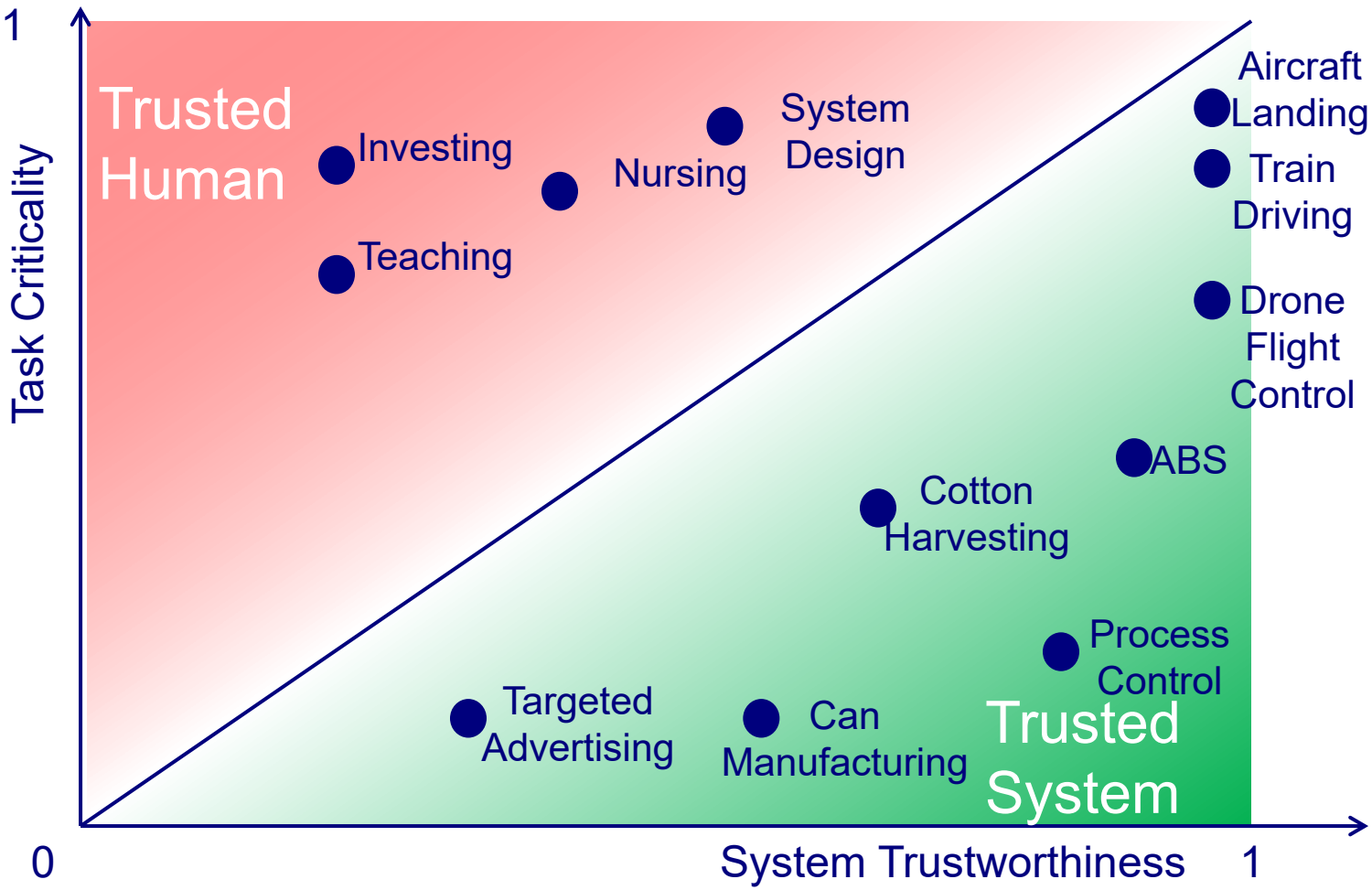
The Automation Frontier –Trustworthiness vs. Criticality



How we decide whether a System can be trusted for performing a Task?

- System Trustworthiness: the system will behave as expected despite any kind of mishaps e.g. resilience to errors, failures, attacks – subsumes functional correctness.
- Task Criticality: characterizes the severity of the impact of an error in the fulfilment of the task e.g. driving a car, operating on a patient, nuclear plant control.

The Automation Frontier – Automated vs.Non-automated



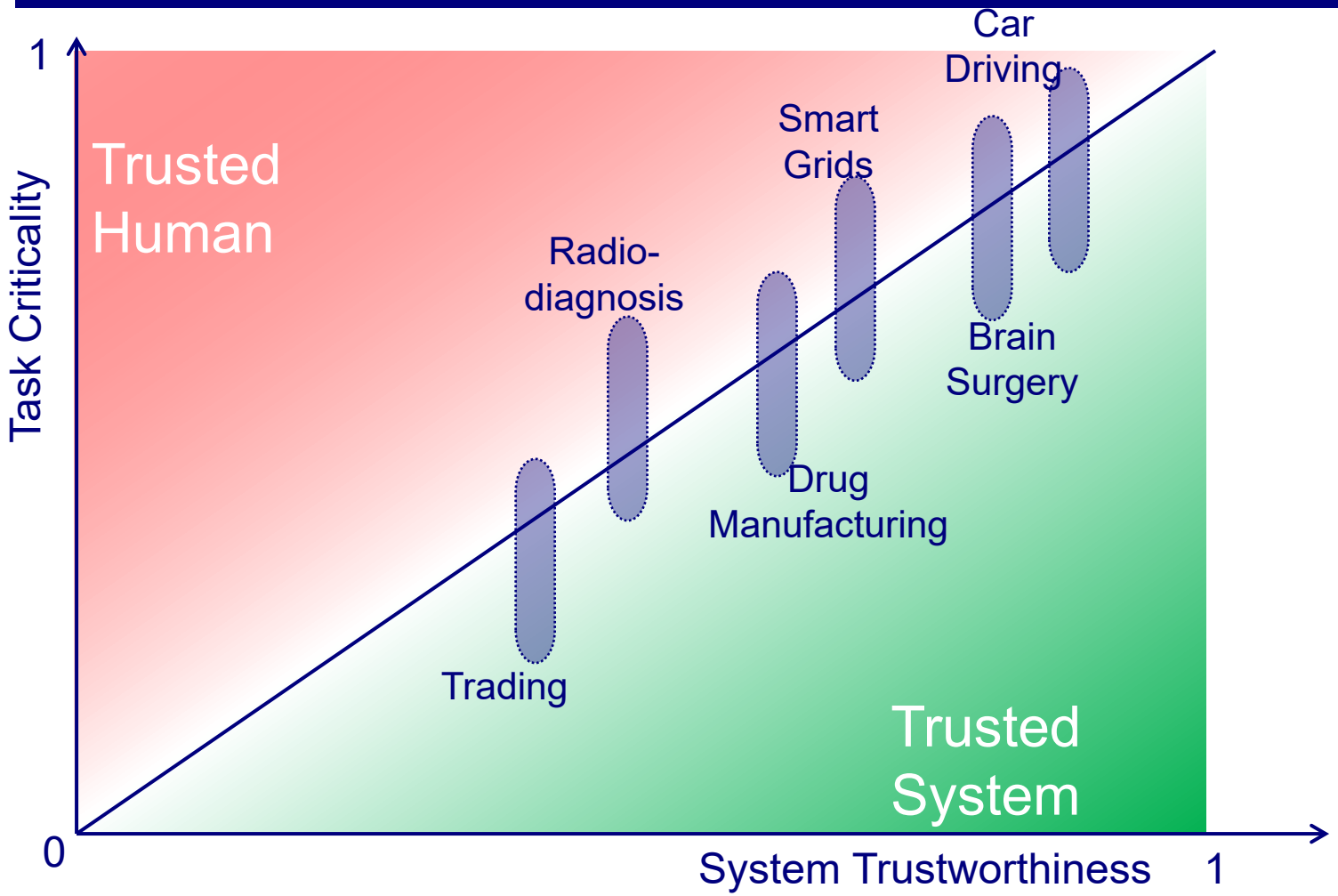
Automated systems: static decision process and/or small impact of failures.

Non-automated systems: require good situation awareness and multiple goal management.

The Automation Frontier – Autonomous Systems

	SAE AUTONOMY LEVELS
Level 0	No automation
Level 1	Driver assistance required (“hands on”)
	The driver still needs to maintain full situational awareness and control of the vehicle e.g. cruise control.
Level 2	Partial automation options available(“hands off”)
	Autopilot manages both speed and steering under certain conditions, e.g. highway driving.
Level 3	Conditional Automation(“eyes off”)
	The car, rather than the driver, takes over actively monitoring the environment when the system is engaged. However, human drivers must be prepared to respond to a "request to intervene”
Level 4	High automation (“mind off”)
	Self driving is supported only in limited areas (geofenced) or under special circumstances, like traffic jams
Level 5	Full automation (“steering wheel optional”)
	No human intervention is required e.g. a robotic taxi

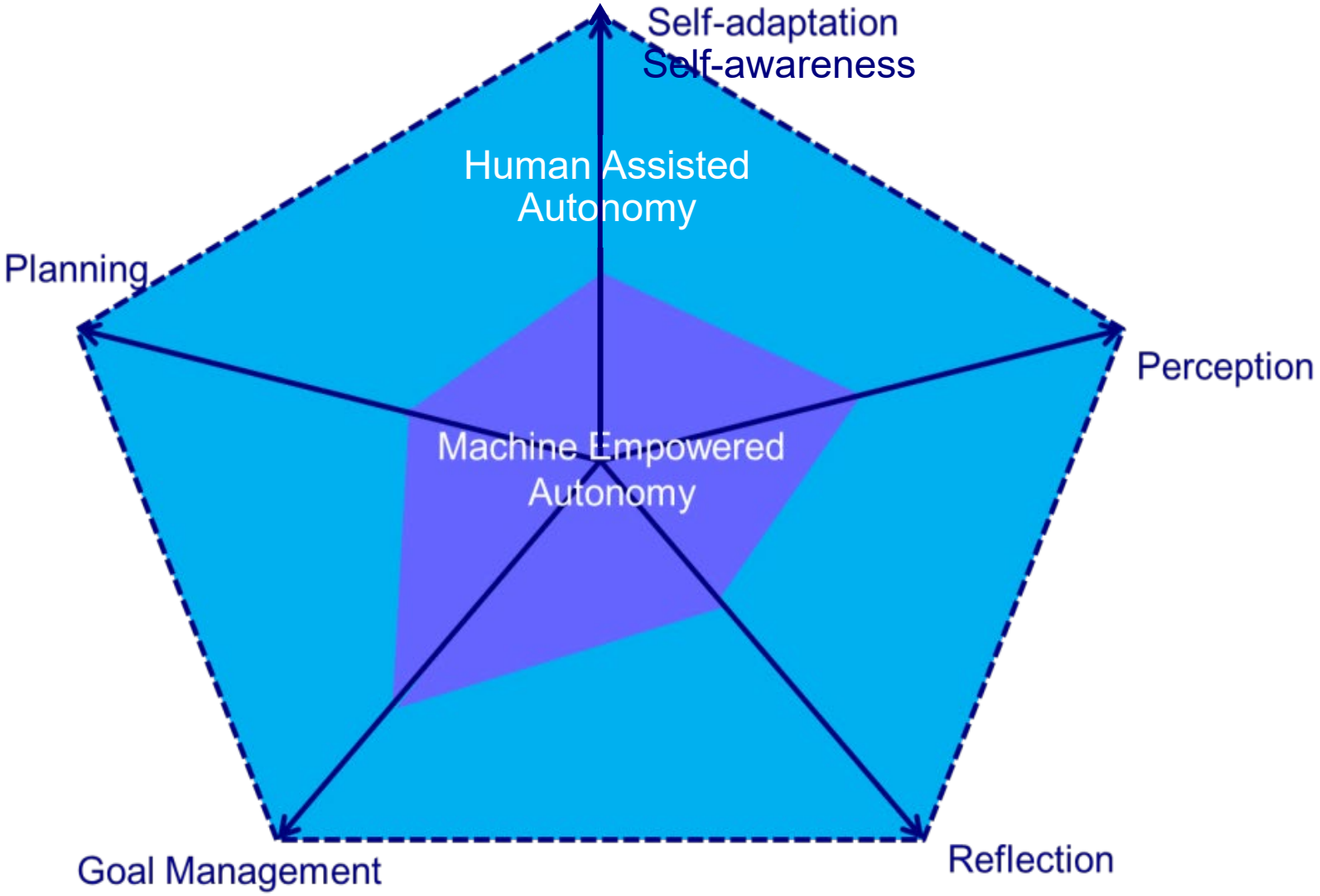
The Automation Frontier – Symbiotic Autonomy



Choose the appropriate autonomy level for which harmonious collaboration between humans and machines can be ensured by protocols enabling

- a human agent to override the machine's decision(s);
- a machine to proactively solicit human agent's intervention.

The Automation Frontier – Symbiotic Autonomy



Problem: Find a division of work that makes the best of the collaboration between human and machine e.g. tele-operated autonomous vehicles

The Automation Frontier – The Role of Institutions

Trustworthiness has a subjective dimension shaped by institutions.

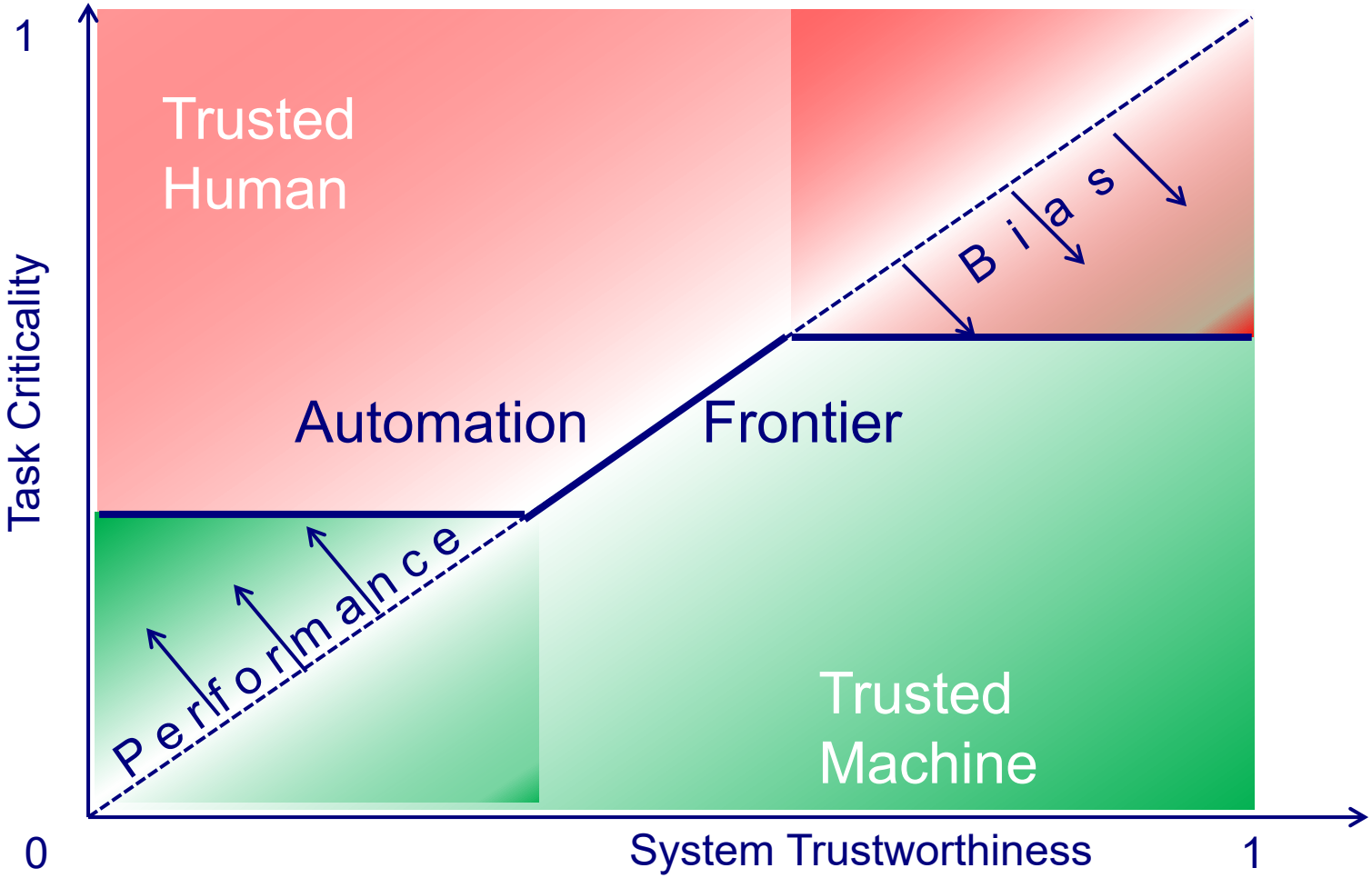
Galileo is WRONG!!

100+ years after

Galileo is RIGHT!!

- Institutions elaborate public perceptions about what is TRUE, RIGHT, SAFE, etc…
- In modern societies, independent institutions guarantee trustworthiness of technical infrastructure and common services based on standards and regulations such as. FDA, FAA, NHTSA, in the USA.
- Note that critical systems standards enforce rigorous design techniques from toasters to bridges and aircraft.
  Such standards are not applicable to ML systems.

The Automation Frontier – Other Shaping Factors



- Performance: for low criticality, trade quality of service for performance e.g. internet bots that fetch, analyze and file information from web servers
- Bias: public opinion is more unforgiving for critical system failures than for human errors e.g. accidents by self-driving car vs. accidents by human drivers

- Autonomous Systems
  - The concept of autonomy
  - The Automation Frontier
- Knowledge Truthfulness
- Design for Trustworthiness and Performance
  - Complexity Issues
  - "Hybrid" design flows
  - Validation
- Discussion

OVERVIEW

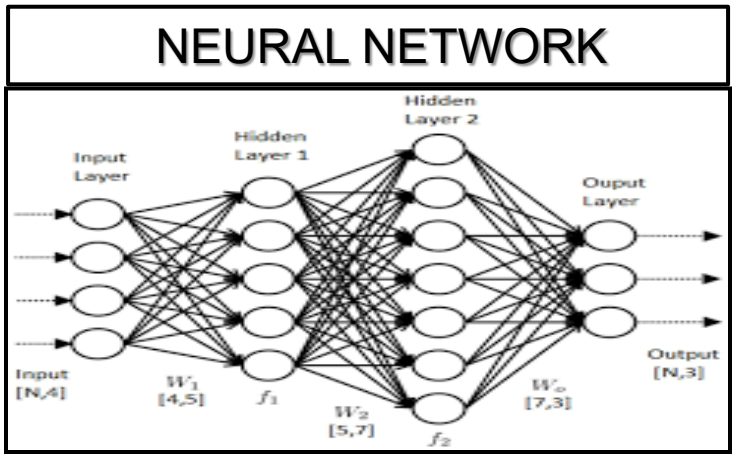
Knowledge Truthfulness – An Interesting Analogy

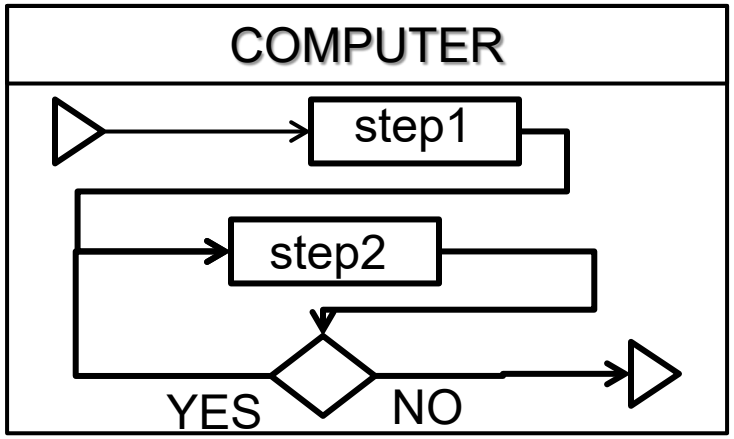
Fast thinking vs. Slow thinking (D. Kahneman's "Thinking Fast and Slow")

System 1: "Fast" Thinking
- Non-conscious – automatic – effortless;
- Without self-awareness or control;
- Handles all kind of empirical implicit knowledge e.g. walking, speaking or playing the piano.

System 2: "Slow" Thinking
- Conscious – controlled– effortful;
- With self-awareness and control;
- Is the source of any reasoned knowledge e.g. mathematical, scientific, technical.

Neural Networks vs. Conventional Computers

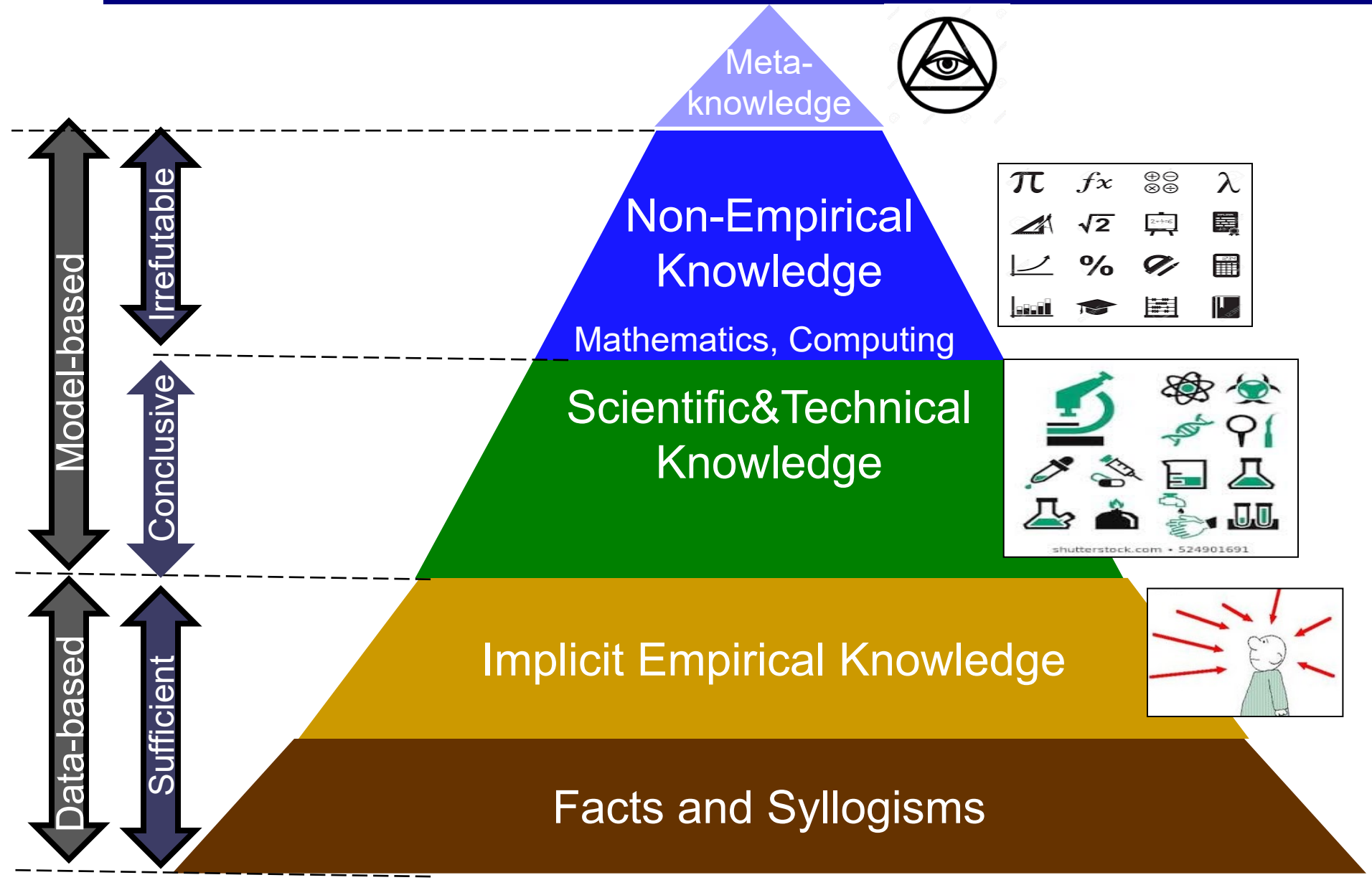




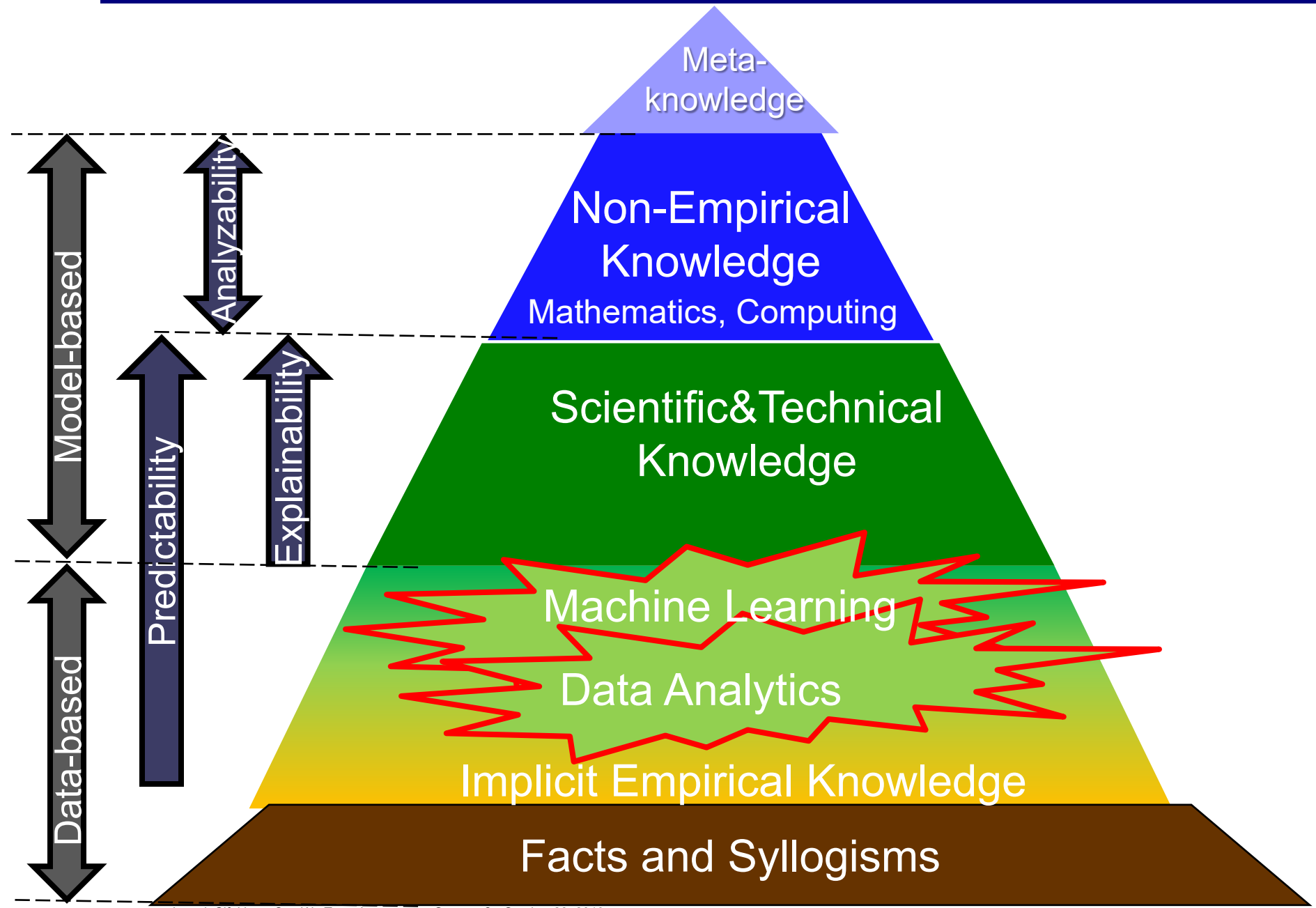
- Generate empirical knowledge after training (Data-based knowledge).
- Distinguish "cats from dogs" exactly as kids do – Cannot be verified!

- Execute algorithms (Model-based knowledge).
- Deal with explicitly formalized knowledge – Can be verified!

Knowledge Truthfulness – The Knowledge Hierarchy



Knowledge Truthfulness – The Knowledge Hierarchy (After)



Knowledge Truthfulness – Scientific vs. ML-generated

1. EXPERIMENT

F1,F2,…Fn

F
(Force)

m
(mass)

a
(Acceleration)

a1,a2,…an

2. LEARNING

Galileo

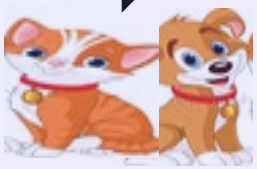
3. EXPLANATION

F = m a
(model)

I1,I2,…In

Image

c1,c2 …cn



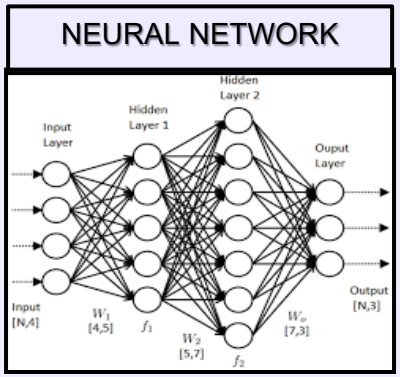


Image
?
{Cat,Dog}

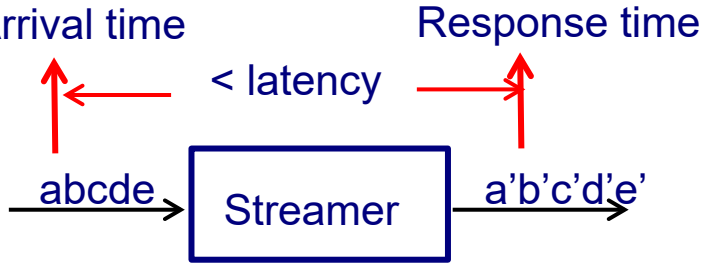
- Autonomous Systems
  - The concept of autonomy
  - The Automation Frontier
- Knowledge Truthfulness
- Design for Trustworthiness and Performance
  - Complexity Issues
  - "Hybrid" design flows
  - Validation
- Discussion

OVERVIEW

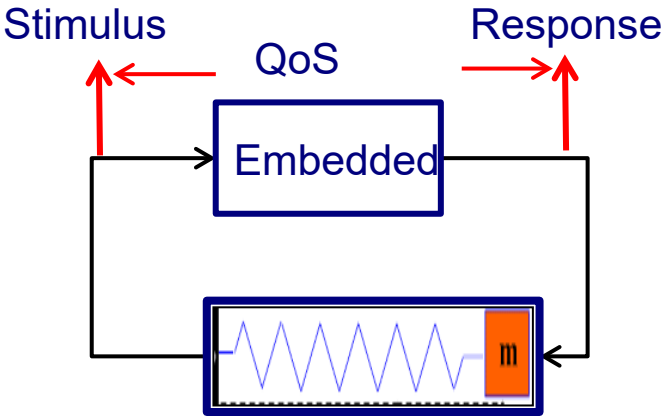
Complexity Issues – Components: Reactive Complexity

Reactive complexity characterizes the intricacy of the interaction between an agent and its environment. It is independent from space complexity or time complexity measuring the quantity of computational resources needed by the agent.

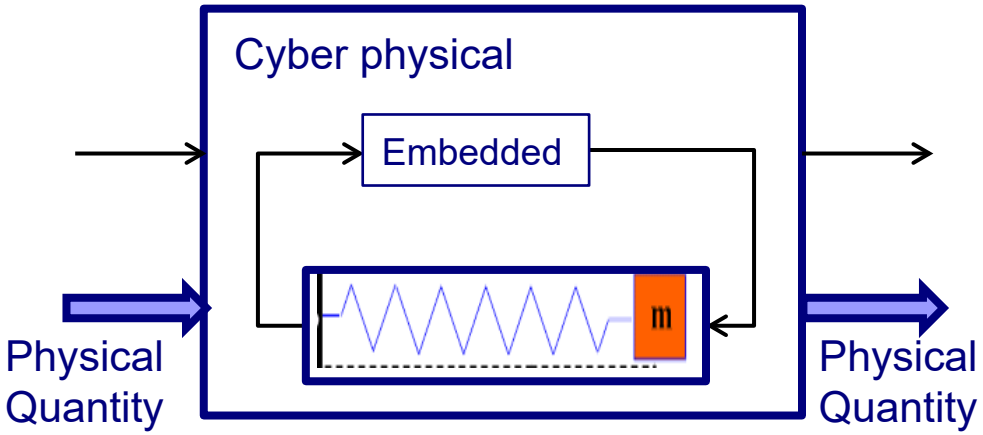
Transformational agent e.g.
Intelligent Personal Assistant



Streaming Agent
e.g. Encoder, Signal processor



Embedded Agent
e.g. Flight controller

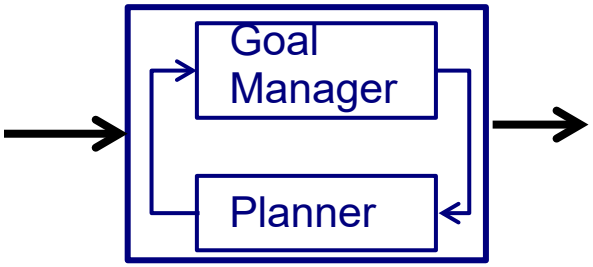


Cyber physical agent
e.g. Self-driving car

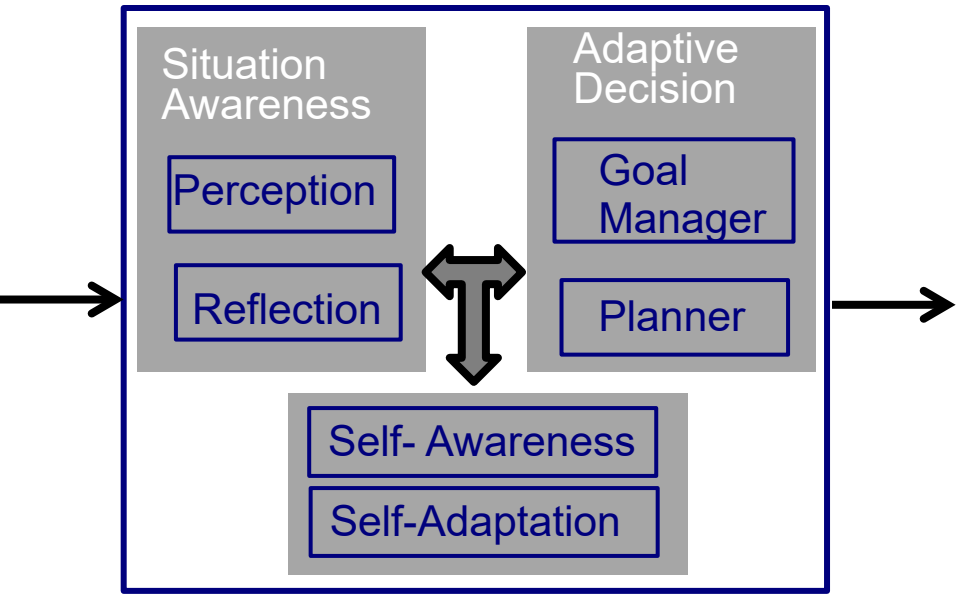
Complexity Issues – Components: Autonomic Complexity

Autonomic complexity: characterizes is the intricacy of the component's task to achieve a set of coordinated goals by its own means adapting to unpredictable environment.

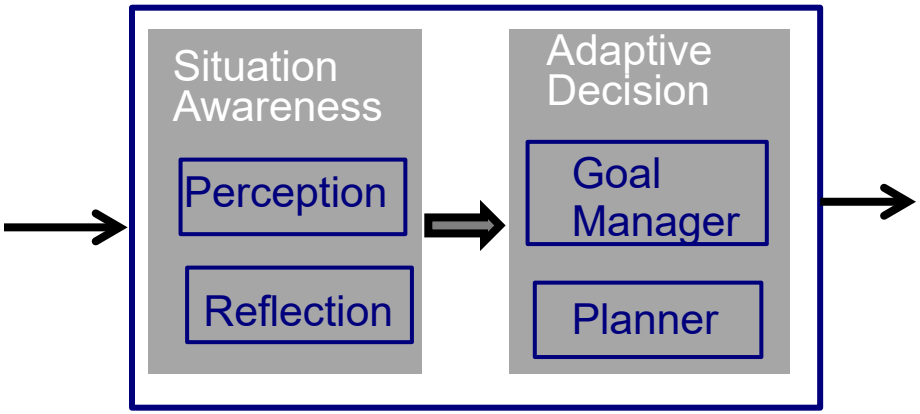
Automated Systems e.g.
Thermostat, Flight Controller



Adaptive Systems e.g.
QoS control, Gaming systems



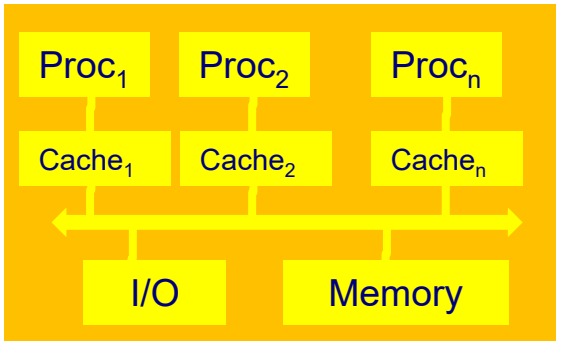
Self-learning Autonomous Agent e.g.
Autonomous Explorer



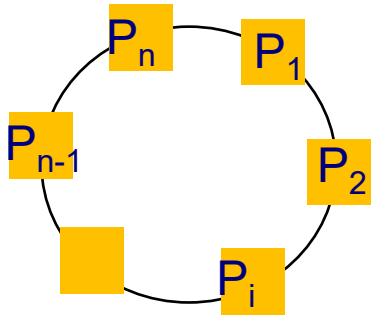
Autonomous Agent e.g.
Autonomous Vehicles

Complexity Issues – Architectural Complexity

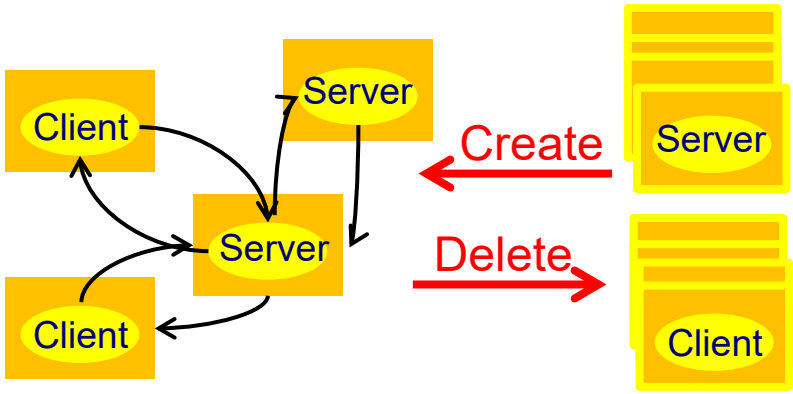
Space and time dynamism of component coordination



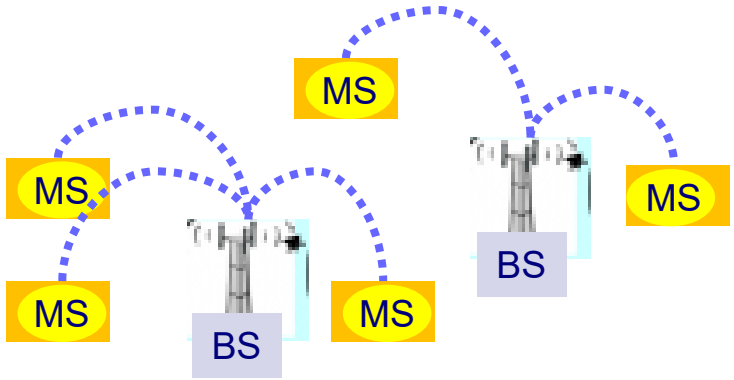
Static Architecture:
Multiprocessor System



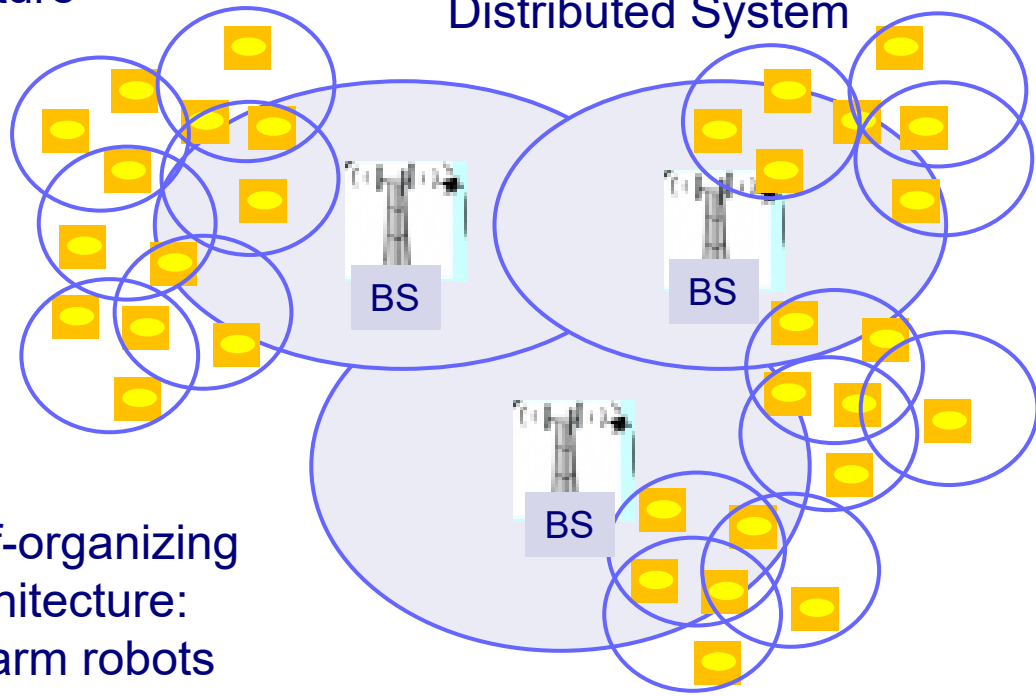
Parametric Architecture:
Ring Architecture



Dynamic Architecture:
Distributed System



Mobile Architecture:
Mobile phones

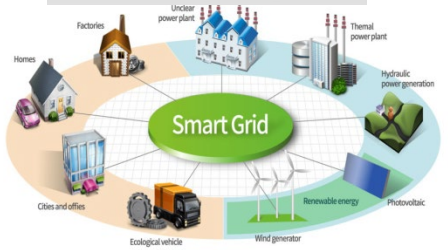


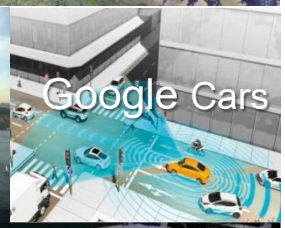
Self-organizing
Architecture:
Swarm robots

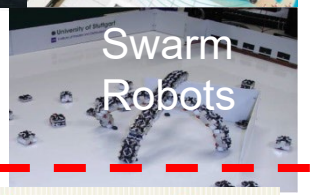
Complexity Issues: Reactive × Architectural Complexity

Agricultural Robot

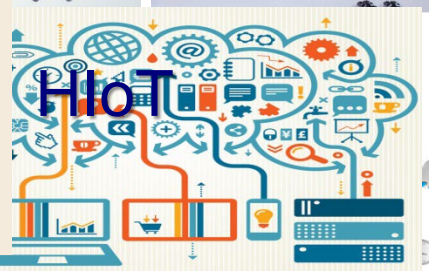
Active safety

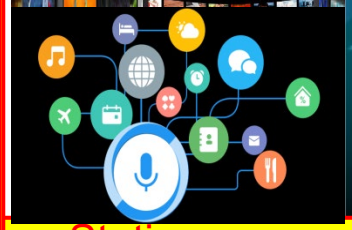






Pacemaker





AGENT REACTIVE COMPLEXITY

CyberPhy

Embedded

Streaming

Transformational

SYSTEMS

SERVICES

Static parametric dynamic mobile self-organizing

ARCHITECTURAL COMPLEXITY

Complexity Issues: Reactive × Architectural Complexity

AGENT REACTIVE COMPLEXITY

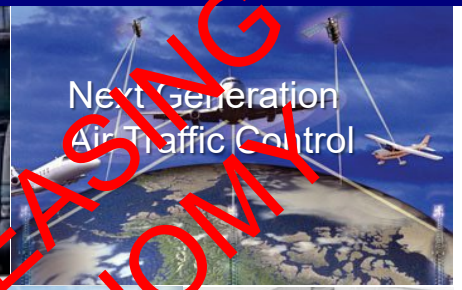
CyberPhy

Embedded

Streaming

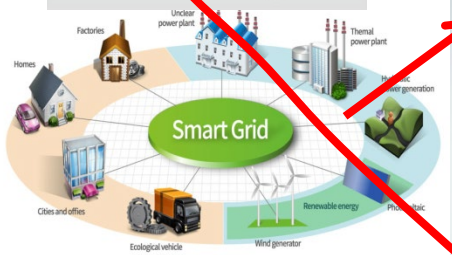
Transformational

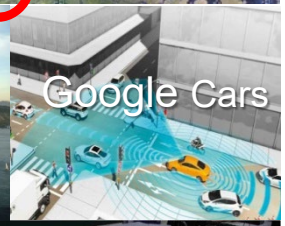
Agricultural Robot

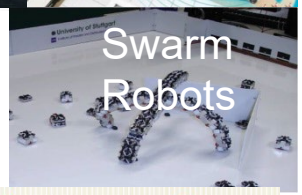


INCREASING AUTONOMY

Active safety







Pacemaker

Connected Medical Devices

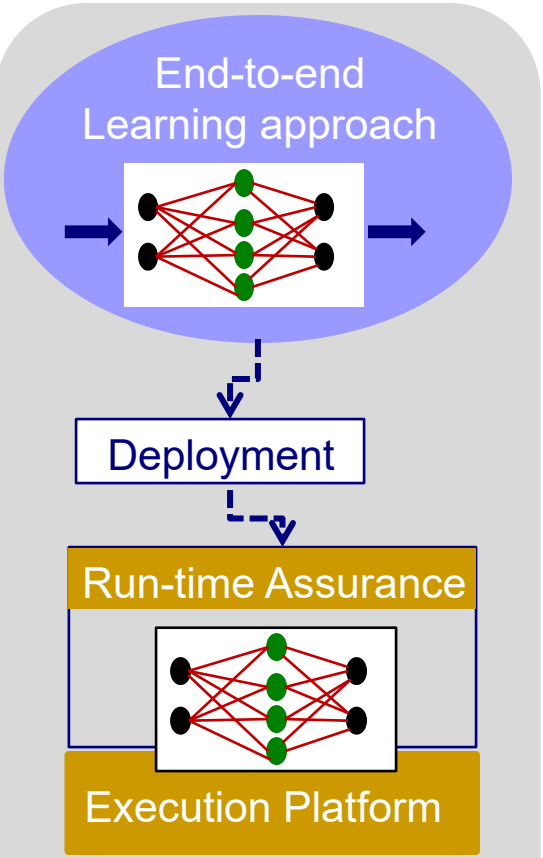
Static parametric dynamic mobile self-organizing

ARCHITECTURAL COMPLEXITY

- Autonomous Systems
  - The concept of autonomy
  - The Automation Frontier
- Knowledge Truthfulness
- Design for Trustworthiness and Performance
  - Complexity Issues
  - "Hybrid" design flows
  - Validation
- Discussion

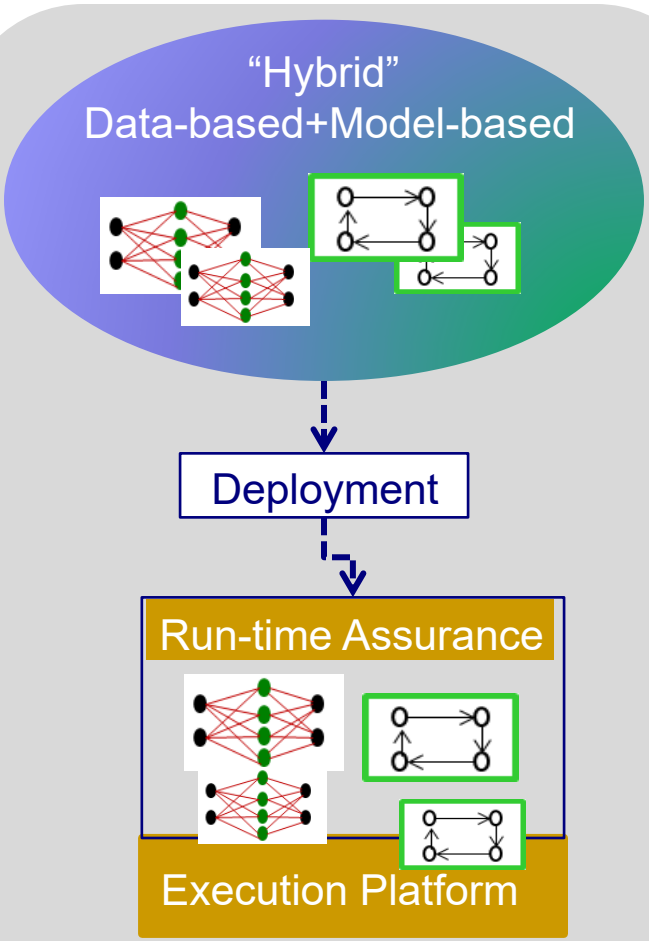
OVERVIEW

“Hybrid” Design Flows – The Principle

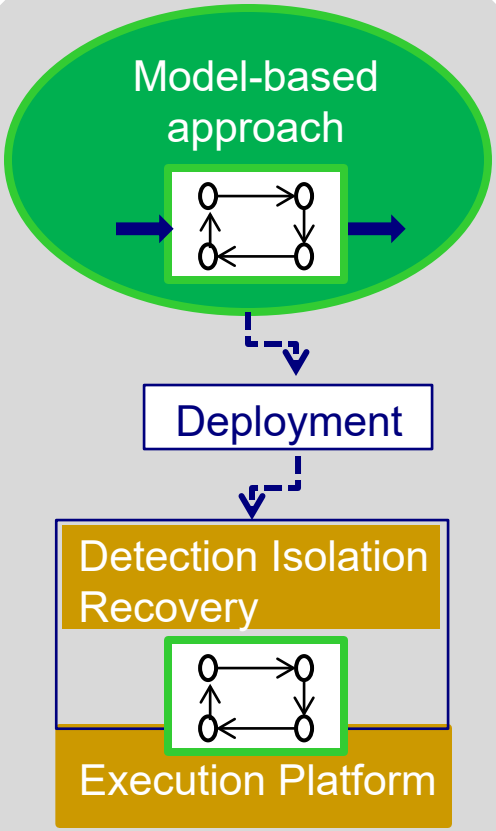


ML-ENABLED AGENT

(PERFORMANCE)



HYBRID AUTONOMOUS AGENT
(TRADE OFF: PERFORMANCE/
TRUSTWORTHINESS)

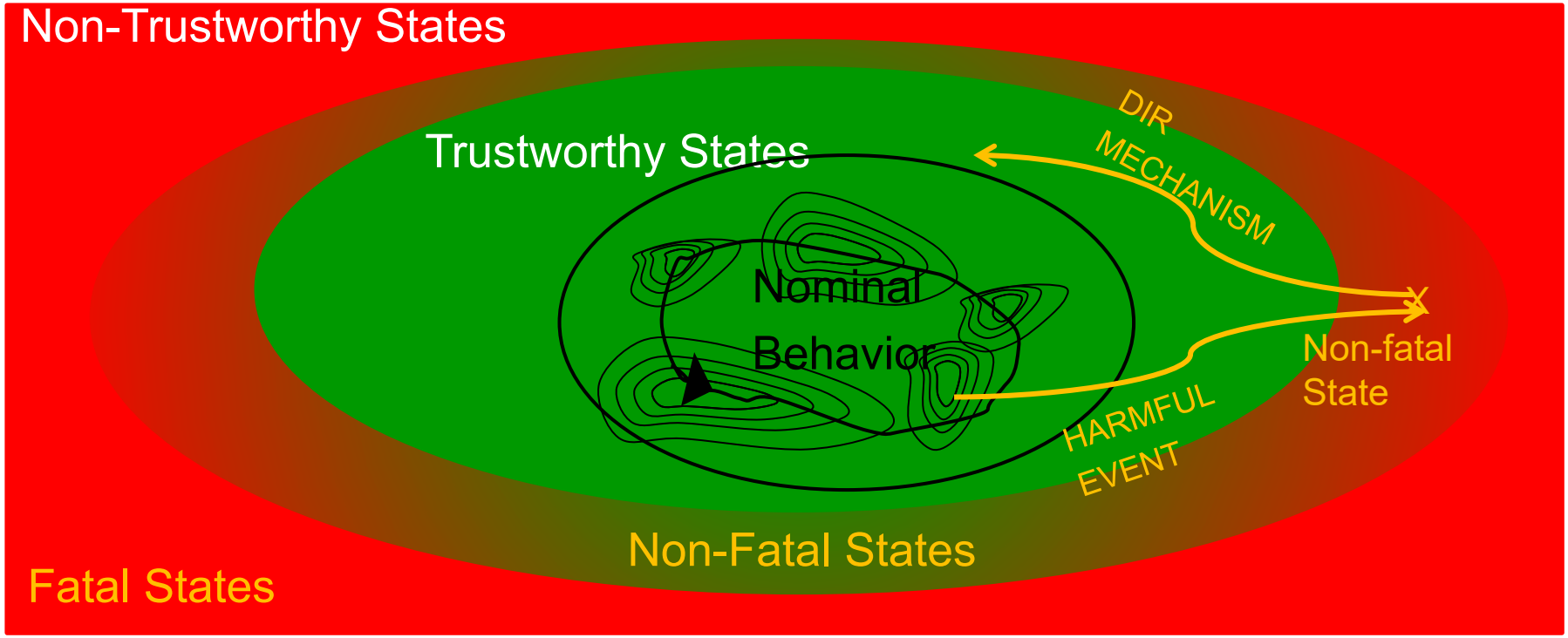


AUTOMATED AGENT

(TRUSTWORTHINESS)

Hybrid Design Flows – Model-based Trustworthiness

- Critical systems engineering ensures trustworthiness at design time by applying
  - risk analysis: identifies more or less exhaustively all kind of harmful events.
  - fault-tolerance techniques: harmful events lead to non-fatal states.
  - DIR mechanisms: bring system from non-fatal states to trustworthy states.



- Model-based approaches cannot be directly applied to autonomous systems:
  - Overwhelming environment complexity and lack of predictability;
  - Use of "black-box" ML-enabled components.

Hybrid Design Flows – Model-based Trustworthiness

1	Vehicle Failure
2	Control Loss With Prior Vehicle Action
3	Control Loss Without Prior Vehicle Action
4	Running Red Light
5	Running Stop Sign
6	Road Edge Departure With Prior Vehicle Maneuver
7	Road Edge Departure Without Prior Vehicle Maneuver
8	Road Edge Departure While Backing Up
9	Animal Crash With Prior Vehicle Maneuver
10	Animal Crash Without Prior Vehicle Maneuver
11	Pedestrian Crash With Prior Vehicle Maneuver
12	Pedestrian Crash Without Prior Vehicle Maneuver
13	Pedalcyclist Crash With Prior Vehicle Maneuver
14	Pedalcyclist Crash Without Prior Vehicle Maneuver
15	Backing Up Into Another Vehicle
16	Vehicle(s) Turning – Same Direction
17	Vehicle(s) Parking – Same Direction
18	Vehicle(s) Changing Lanes – Same Direction
19	Vehicle(s) Drifting – Same Direction
20	Vehicle(s) Making a Maneuver – Opposite Direction
23	Lead Vehicle Accelerating
24	Lead Vehicle Moving at Lower Constant Speed
25	Lead Vehicle Decelerating
26	Lead Vehicle Stopped
27	Left Turn Across Path From Opposite Directions at Signalized Junctions
28	Vehicle Turning Right at Signalized Junctions
29	Left Turn Across Path From Opposite Directions at Non-Signalized Junctions
30	Straight Crossing Paths at Non-Signalized Junctions
31	Vehicle(s) Turning at Non-Signalized Junctions
32	Evasive Action With Prior Vehicle Maneuver
33	Evasive Action Without Prior Vehicle Maneuver
34	Non-Collision Incident
35	Object Crash With Prior Vehicle Maneuver
36	Object Crash Without Prior Vehicle Maneuver
37	Other

Pre-crash failure typology covering 99.4% of light-vehicle crashes for 5,942,000 cases.
Source: Pre-Crash Scenario Typology for Crash Avoidance Research, DOT HS 810 767, April 2017.

Run-time assurance techniques: replace DIR at design time by run-time monitoring

Hybrid Design Flows – Model-based Guarantees

Mobileye's Responsibility-Sensitive Safety (*): Compute lower bounds of the distance between two cars that guarantee safety.

(*) "*On a Formal Model of Safe and Scalable Self-driving Cars*" Shai Shalev-Shwartz, Shaked Shammah, Amnon Shashua, Mobileye, 2017

Safe Distance Formula

$$d_{\min} = L + T_f\,[v_r - v_f + \rho\,(a_a + a_b)] - \frac{\rho^2 a_b}{2} + \frac{(T_r - T_f)(v_r + \rho\, a_a - (T_f - \rho)a_b)}{2}$$

- L is the average length of the vehicles
- ρ is the response time of the rear vehicle
- v_r, v_f are the velocities of the rear/front vehicles
- a_a, a_b are the maximal acceleration/braking of the vehicles
- T_f is the time for the front car to reach a full stop if it would apply maximal braking
- T_r is the time for the rear car to reach a full stop if it would apply maximal acceleration during the response time, and from there on maximal braking

Beware! Safety cannot be dissociated from performance e.g. on a two-lane road the car should an overtaking car on the left lane should move safely as fast as possible

Hybrid Design Flows – Model-based Adaptive Decision

- Premises for "hybrid" autopilot design:
  - Situation awareness is ML-enabled while adaptive decision is model-based.
  - The perception function can recognize a well-defined and "complete set" of environment configurations corresponding to "driving modes" each mode requiring a specific maneuver (*)

- The decision process is hierarchically structured:
  - Level 1: Anti-collision system (Acceleration) and Trajectory Tracking Control System (Steering angle)
  - Level 2: Maneuver Protocols (driving modes) e.g. Overtaking, Platooning, Roundabout Movement, Parking, etc.
  - Level 3: Environment model and analysis including Safety envelope computation, Trajectory computation, Driving mode selection.
  - Level 4: Itinerary Goals and Planning

(*) 19 Tactical and Operational Maneuvers enumerated in NHTSA Report "A Framework for Automated Driving System Testable Cases and Scenarios", Sept 2018.

- Autonomous Systems
  - The concept of autonomy
  - The Automation Frontier
- Knowledge Truthfulness
- Design for Trustworthiness and Performance
  - Complexity Issues
  - "Hybrid" design flows
  - Validation
- Discussion

OVERVIEW

Validation – Modeling and Simulation

Simulation is of paramount importance for validation – whatever approach is taken - and covers a large variety of aspects from purely technical to theoretical ones.

1. Realism: agent behavior and environment look real in a way that is accurate or true to life.

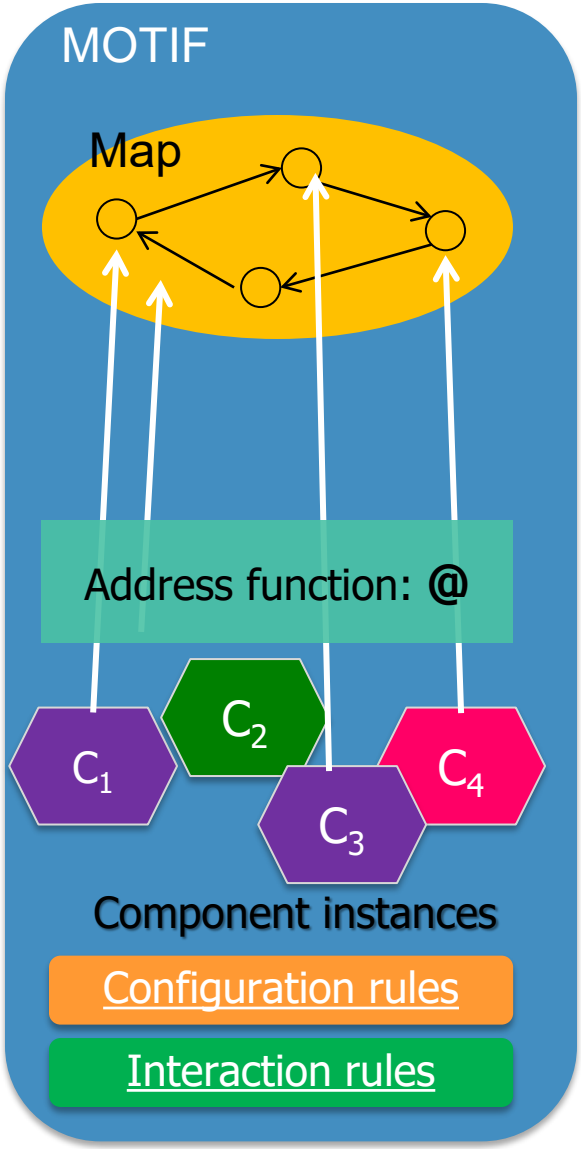
2. Semantic awareness: the simulated system dynamics is rooted in transition system semantics.
   - Notion of state allowing controllability and repeatability of experiments.
   - Scenarios to explore/detect corner cases and high risk situations
   - Notion of coverage measuring the degree to which relevant system configurations have been explored.

3. Multiscale multigrain modeling and simulation
   - Theory: cyber physical systems modeling; correlation between scales …
   - Practice: run-time infrastructure federating simulation engines e.g. HLA, FMI

Most industrial simulation systems lack semantic awareness e.g. rely on game engines or pre-built software.

What is the value of results reported by Waymo: 27 000 cars running 24/7, 10 million miles simulated per day, >7 Billion miles of simulation?.

Validation – Modeling and Simulation: DR-BIP



DR-BIP (Dynamic Reconfigurable BIP)

- A system is a set of (architecture) motifs

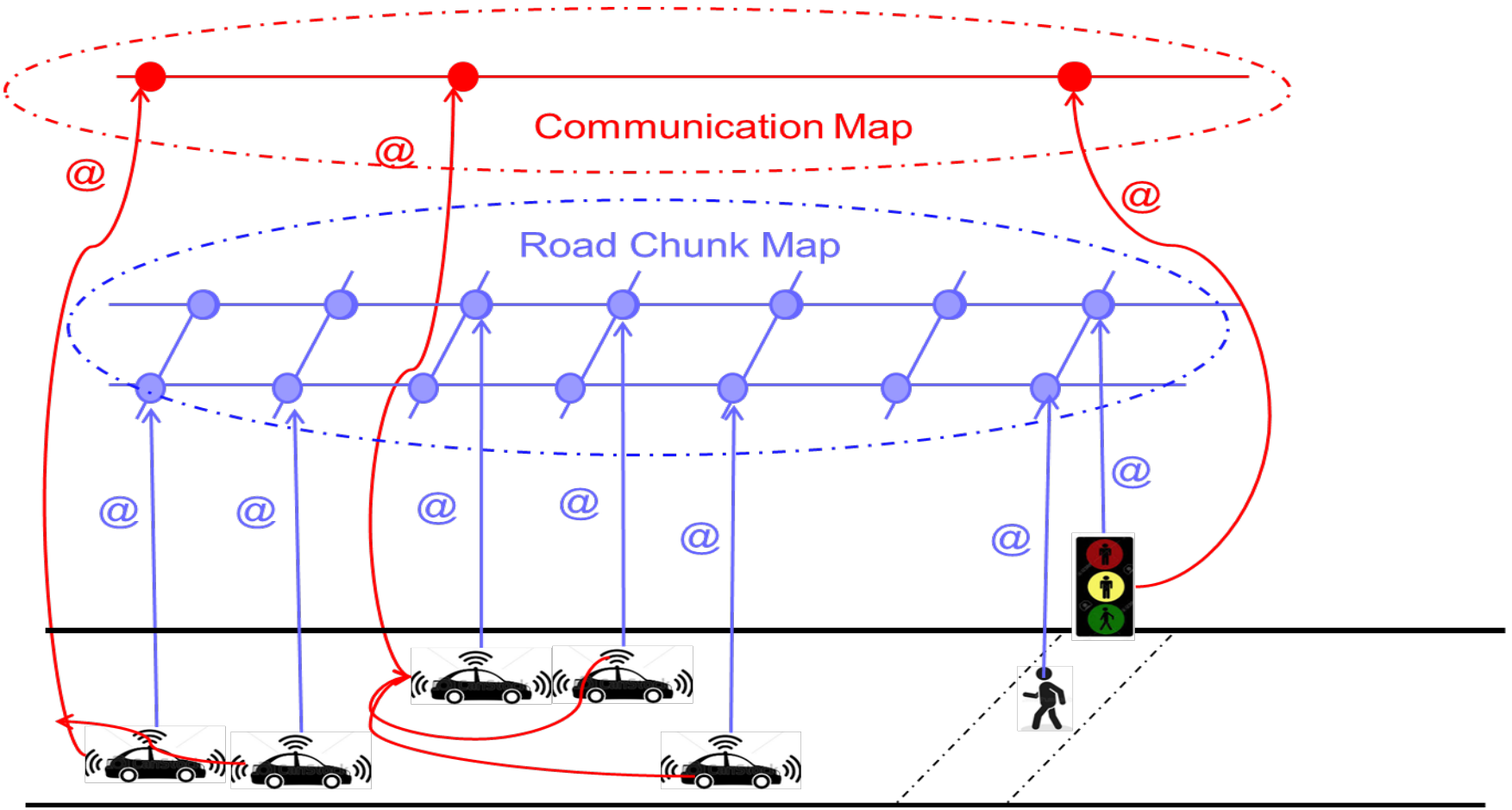
- A motif is a coordination mode consisting of
  - A set of components, instances of types of agents or objects
  - A map that is a graph (N,E) used to describe relations between components e.g. geographical, organizational, etc.
  - An address function @ mapping components into nodes of the map

  - Interaction rules: define interactions (atomic multiparty synchronization) between components

  - Configuration rules:
    - Mobility of components (change of @)
    - Creation/deletion of components
    - Dynamic change of the map

The meaning of systems models is defined using operational semantics

Validation – Modeling and Simulation: DR-BIP



Interaction rule:
for all a,a':vehicle, if [dist(@(a),@(a'))<l] then exchange(a.speed,a'.speed).

Mobility rule :
for all a:vehicle if @(a)=n and $@^{-1}(n+1)$=empty then @(a):=n+1.

Validation – Verification

- As a rule, ML systems cannot be formally verified:
  - they are not developed based on formal goals e.g. specifying how a dog looks different from a cat;
  - they remain mostly “black boxes” although recent work on “Interpretable AI” allows characterization of I/O behavior of particular classes of DNN e.g. ReLU.

- Formal verification suffers (well-known) limitations:
  - is applicable when goals and requirements can be formalized;
  - can be automated for “monolithic” models representing the global system behavior (state explosion lack of compositional verification techniques);
  - is practically limited to static architectures – while autonomous systems are naturally dynamic and reconfigurable!
  - Is not enough!
    Autonomy is about controller synthesis under both safety and optimization constraints.

Validation – Formalizing Requirements

Formalization of requirements for autonomous systems is extremely hard e.g. "behavioral competencies" for self-driving cars (California PATH)

1. 1. Detect and Respond to Speed Limit Changes and Speed Advisories
2. [illegible]
3. Perform Low-Speed Merge
4. Move Out of the Travel Lane and Park (e.g., to the Shoulder for Minimal Risk)
5. [illegible]
6. 6. Detect Passing and No Passing Zones and Perform Passing Maneuvers
7. [illegible]
8. Detect and Respond to Stopped Vehicles
9. Detect and Respond to Lane Changes
10. Detect and Respond to Static Obstacles in the Path of the Vehicle
11. Detect Traffic Signals and Stop/Yield Signs
12. Respond to Traffic Signals and Stop/Yield Signs
13. 13. Navigate Intersections and Perform Turns
14. [illegible]
15. Navigate a Parking Lot and Locate Spaces
16. Detect and Respond to Access Restrictions (One-Way, No Turn, Ramps, etc.)
17. [illegible]
18. 18. Make Appropriate Right-of-Way Decisions
19. [illegible]
20. Follow Police/First Responder Controlling Traffic (Overriding or Acting as Traffic Control Device)
21. Follow Construction Zone Workers Controlling Traffic Patterns (Slow/Stop Sign Holders).
22. Respond to Citizens Directing Traffic After a Crash
23. Detect and Respond to Temporary Traffic Control Devices
24. Detect and Respond to Emergency Vehicles
25. Yield for Law Enforcement, EMT, Fire, and Other Emergency Vehicles at Intersections, Junctions, and Other Traffic Controlled Situations
26. Yield to Pedestrians and Bicyclists at Intersections and Crosswalks
27. [illegible]
28. 28. Detect/Respond to Detours and/or Other Temporary Changes in Traffic Patterns

- Autonomous Systems
  - The concept of autonomy
  - The Automation Frontier
- Knowledge Truthfulness
- Design for Trustworthiness and Performance
  - Complexity Issues
  - "Hybrid" design flows
  - Validation
- Discussion

OVERVIEW

Discussion – What the Future can be Like?

- AV manufacturers revise their ambitions because of technical problems and the erosion of public trust (*).

(*) "*I think both industry and media have been complicit in hyping this and not being open and honest enough about the realities of the technology*."
Jack Weast, vice president, autonomous vehicle standards, Intel, July 2019.

- Go beyond the debate opposing data-based and model-based approaches develop hybrid design flows.

  - agree on "hybrid" design principles relying on architectural decomposition into base functions and their interconnection.
  - seek trade offs between performance and trustworthiness
  - agree on trustworthiness evaluation principles (<u>much more than random testing!</u>) that could be a basis for standards

- Prepare the way for smooth and progressive transition along the different levels toward full autonomy or symbiotic autonomy

Discussion – What the Future can be Like?

❑ Standards and regulations

In the US, state and federal legislation and regulations on autonomous vehicles have been largely permissive with a focus on "getting the technology on the streets".

- There are no explicitly defined criteria for assessing AV trustworthiness
  e.g. specific "thresholds" or "requirements" that AVs must clear.
- Automotive and medical systems are "self-certified" by their manufacturers according to guidelines requiring *sufficient evidence* that the developed system is trustworthy enough – instead of *conclusive evidence* of critical system standards!

Will this "provisional" situation become permanent?

❑ Social awareness and sense of responsibility:

- When machines use knowledge in critical decision processes make sure that it is truthful, unbiased, neutral, fair, etc. (precautionary principle).
- Question motives, objectives and biases of existing systems.
- Crucial question: should we grant the power of decision to autonomous systems without rigorous and strict guarantees on the grounds of a (disputable) performance benefit.

Discussion – What the Future can be Like?

- We are living the beginning of a grand revolution where machines are called to progressively replace humans in their capacity for situation awareness and adaptive decision making.
- This is a first step toward general AI that goes far beyond the objectives of ML-enabled intelligence.
- The role of autonomous systems will depend on choices we make about when we trust them and when we do not.
- Giving ourselves the means to make informed decisions is essential.