AI: Where We Are, Where Are We Going?

Can Machines Save the World?
"The Digital Humanism Fellowship Conference".

Vienna
November 16, 2023

Joseph Sifakis
Verimag Laboratory

AI – Where We Are, Where Are We Going?

At present, there's a great deal of confusion as to the final objective, with opinions divided between two very different positions:

- Some AI research and companies such as OpenAI and DeepMind see AGI, an ill-defined term, as the ultimate goal
  - suggesting that AGI can be achieved through machine learning and its further developments - it's just a matter of time!
  - focusing on building "super-intelligent agents" capable of analyzing large datasets, identifying patterns and efficiently making data-driven decisions in a variety of sectors, from healthcare and finance to transportation and manufacturing.

- Others see the goal of AI as building machines with human-level intelligence, which requires agreement on what human intelligence is and, more importantly, on methods for comparing human and machine intelligence.

  - According to the Oxford dictionary, intelligence is defined as
    "the ability to learn, understand and think in a logical way about things; the ability to do this well"

  - Machines can do impressive things by outperforming humans in the execution of particular tasks, but they cannot surpass them in situational awareness, adaptation to changes in their environment and creative thinking..

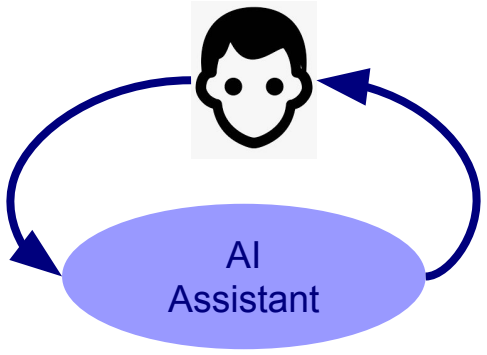
Where We Are? – Three Modes of Use

- Despite the impressive growth in AI we have seen in recent years, culminating in the arrival of generative AI and its application to solving NLP problems that have always remained open, we only have weak AI that
  - gives us only the elements to build intelligent systems but we have no principles and techniques to synthesize them e.g. like we build bridges and buildings.
  - focuses on Intelligent Assistants that interact with a user to provide a service, for example in question-and-answer mode.

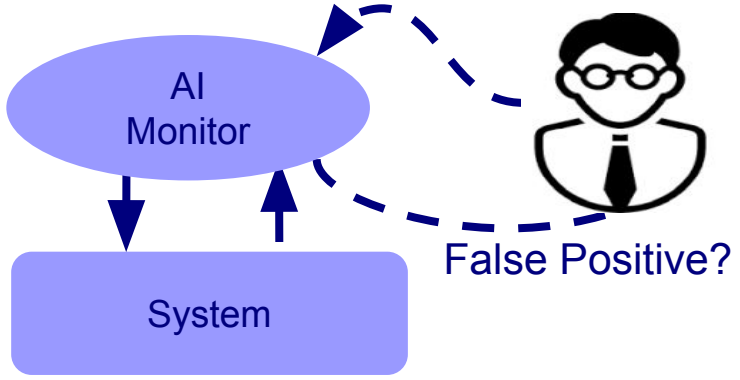
- There are three different ways to use AI systems :

1. Assistants that in interaction with a user, provide a given service;
2. Monitors of a system behavior synthesizing knowledge to detect or predict critical situations;
3. Controllers of a system so that its behavior meets a given set of requirements, e.g. the autopilot of an autonomous car.

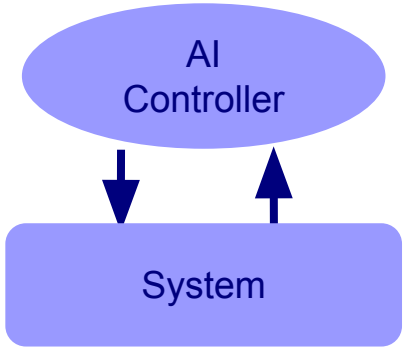
Monitors and Controllers will be by far the most important in the future for building intelligent products and services.



Intelligent Assistant



Monitor for Detection/Prediction



End-to-end Controller
for autonomous behavior

Where We Are? – Toward Autonomous Systems

- Autonomous systems are a bold step toward building systems exhibiting human-level intelligence.
  - They stem from the needs to further automate existing organizations by gradually replacing humans with autonomous agents, as envisioned by the IoT e.g. autonomous cars, smart grids, smart factories, smart farms, autonomous networks.
  - They support a paradigm of intelligent systems that goes beyond machine learning systems, which are often specialized transformational systems
  - They are distributed systems of agents that are often critical and exhibit "broad intelligence" by handling knowledge
    - managing dynamically changing sets of possibly conflicting goals;
    - coping with uncertainty of complex, unpredictable cyber physical environments;
    - harmoniously collaborating with human agents e.g. "symbiotic" autonomy.

- The realization of the autonomy vision is hampered by non explainability AI systems and by difficult systems engineering problems unrelated to agent intelligence - as we have learned from the setbacks of the autonomous car industry, which has had to drastically revise its optimistic forecasts.

- At present, two different technical avenues are unable to meet needs:
  - traditional model-based critical systems engineering, successfully applied to aircraft and production systems, proves to be inadequate.
  - industrial end-to-end AI-enabled solutions that fail to provide the required strong trustworthiness guarantees.

Where We Are? – Guaranteeing Properties of AI Systems

- The extensive use of AI systems - reputed to be "black boxes" - raises questions about their trustworthiness characterized by a set of properties including safety and security.
  - It is impossible to apply verification techniques that are essential to obtain strong trustworthiness guarantees.
  - In particular, AI safety has preoccupied authorities in both the USA and Europe who, despite numerous consultations and legislative efforts, have been unable to draw up a concrete and realistic regulatory framework to date.

- In addition to trustworthy AI, great deal of work is aimed at building AI systems that satisfy human-centric properties
  - "Responsible AI" implies that the development and use of AI meets criteria such as fairness, reliability, safety, privacy and security, inclusiveness, transparency, and accountability, difficult, if not impossible, to assess.
  - “AI alignment" meaning alignment of a conversational agent with human values while we do not even understand how human will emerges and the associated value-based decision-making system works.
  - The properties of *mental attitudes* such as *belief*, *desire* and *intention* are superficially attributed to AI systems.
- But all this work lacks foundation, because it ignores a basic epistemic principle: any claim that a system satisfies a property must be backed up by a rigorous method of validation.

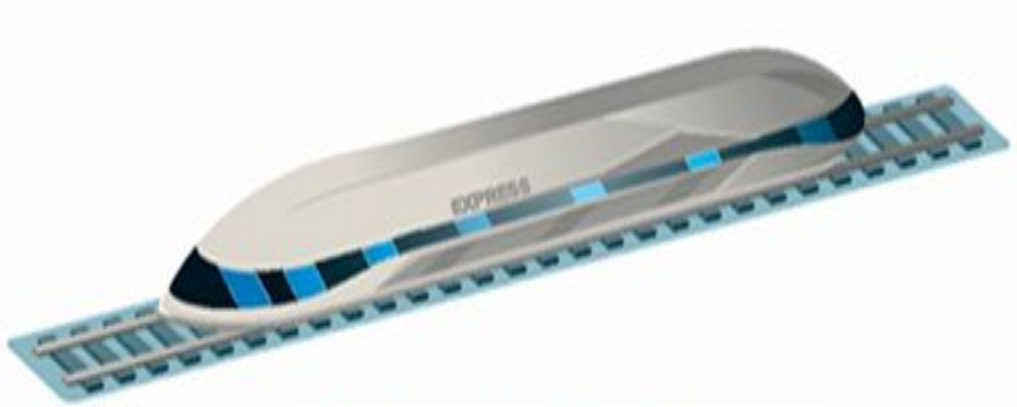
- Can the properties of AI systems be guaranteed in the same way as the properties of traditional digital systems?
  - How traditional systems engineering help us to tackle the problem of guaranteeing the properties of AI systems?
  - Is it possible to transpose existing systems engineering methodologies to AI systems? If so, what are the obstacles?

Overview

- Autonomous Systems
- Validation of Intelligent Systems
- Where Are We Going?

Autonomous Systems – Comparison with Automated Systems

Thermostat



Automated Shuttle

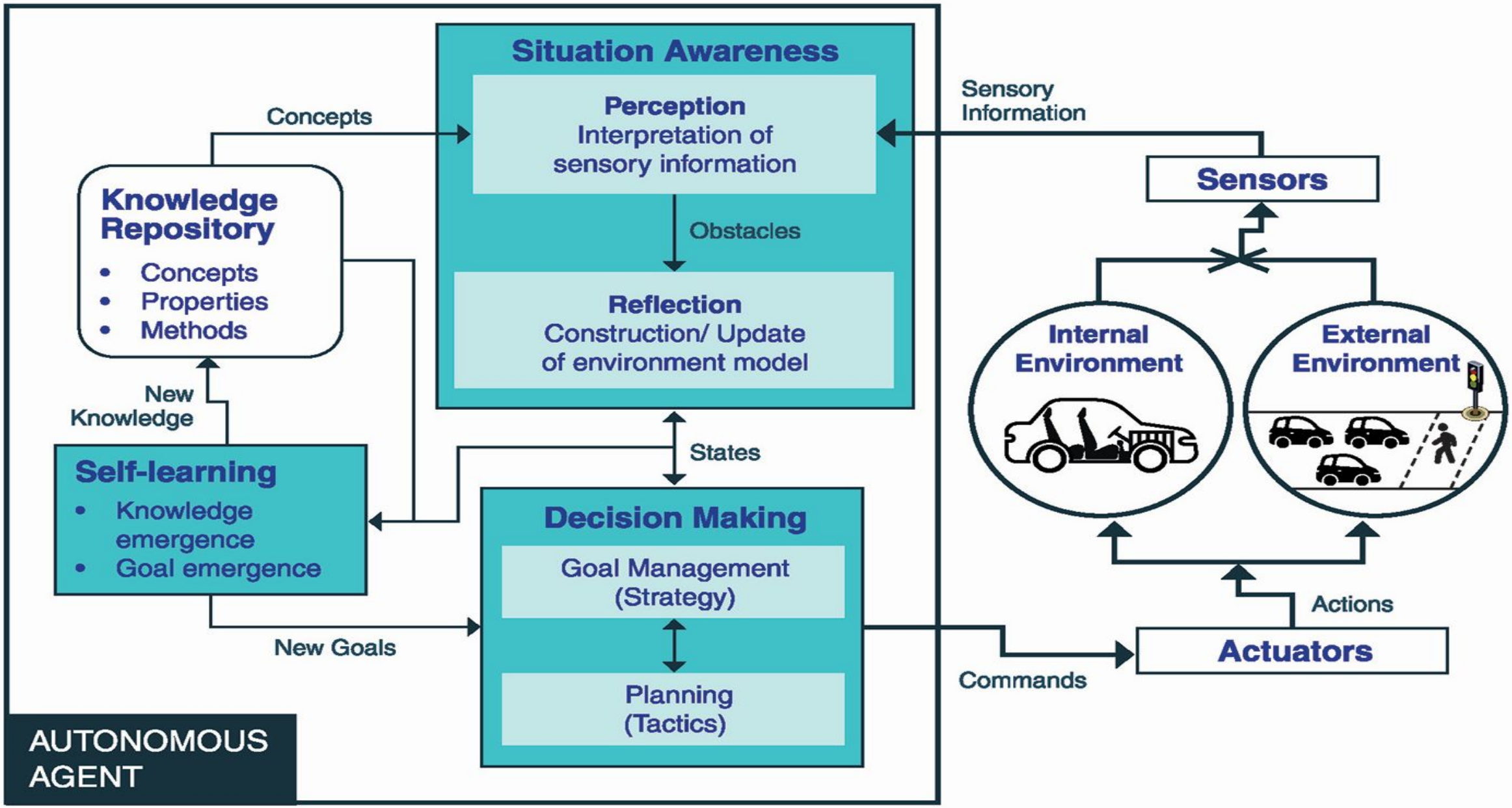
Chess Robots

Football Robots

Autonomous Car

	Thermostat	Shuttle	Chess Robot	Football Robot	Autonomous Car
SITUATION AWARENESS	Temperature (number)	Distance from next stop (number)	Pawns on the board (static image)	Players on the pitch (dynamic image)	Obstacles on the road (dynamic image)
DECISION MAKING	Static goals <18 □ ON >22 □ OFF	Static goals ▪ Stop ▪ Accelerate ▪ Decelerate	Static well-defined goals Dynamic planning of goals	Dynamic multiple goals Dynamic planning of goals	Dynamic multiple conflicting goals Dynamic planning of goals

Autonomous Systems – Autonomous Agent Architecture

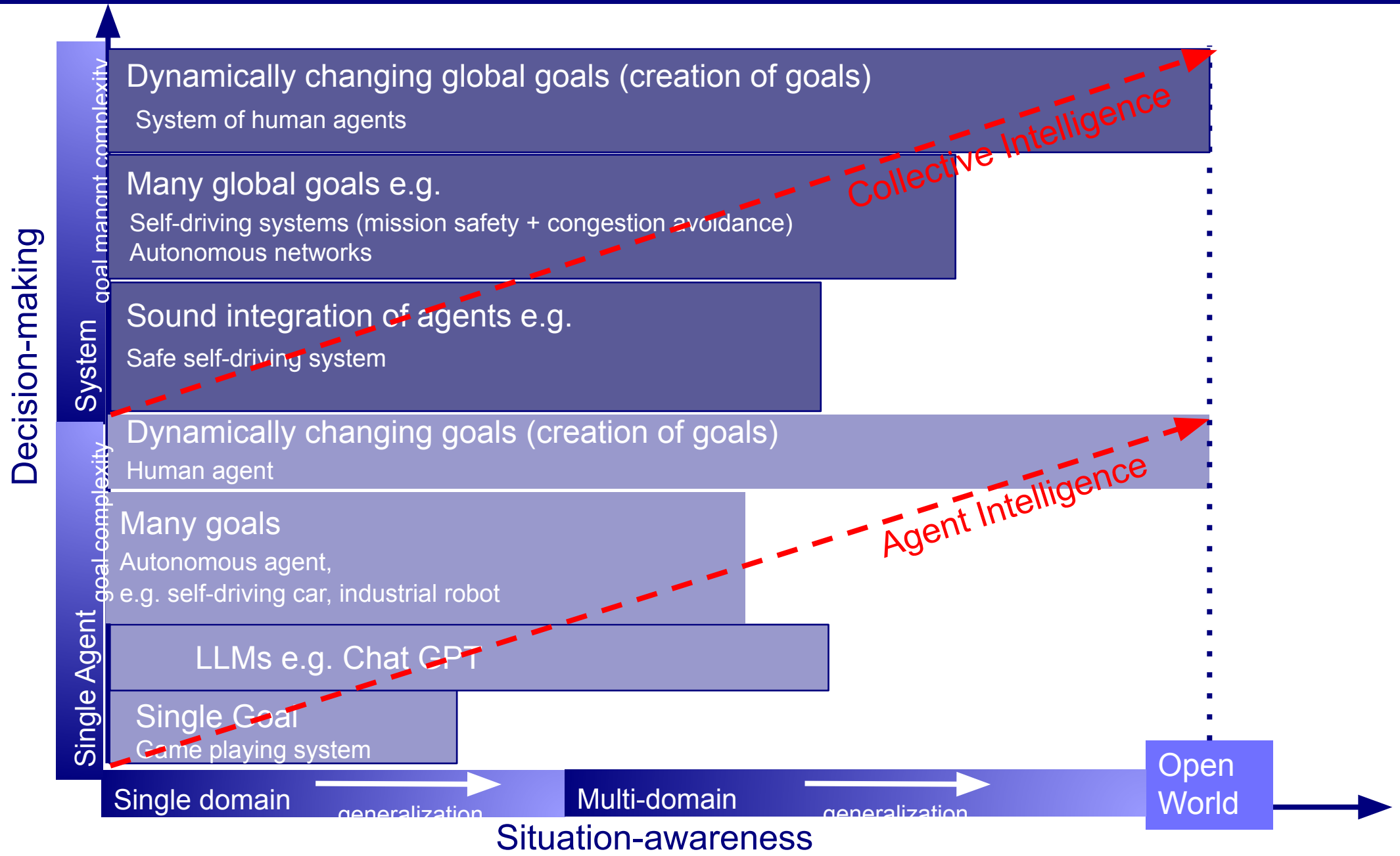


Autonomous Systems – Complexity Issues

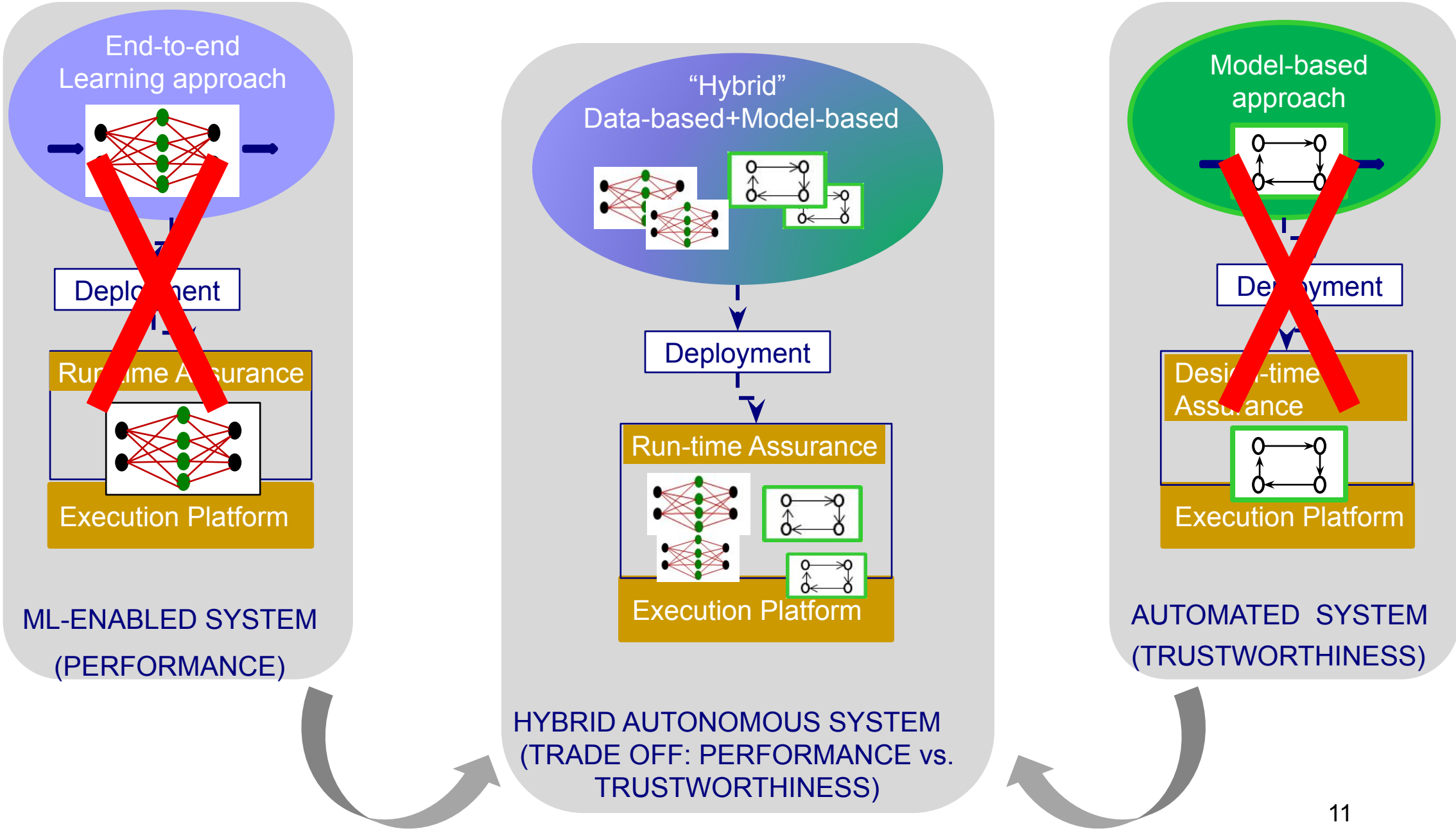
- The construction of autonomous agents is hampered by three different complexities:
  - Complexity of perception due to the difficulty to interpret stimuli (cope with ambiguity, vagueness) and to timely generate corresponding inputs for the agent environment model.
  - Complexity of uncertainty due to situations involving imperfect or unknown information implying lack of predictability about the environment such as dynamic change caused by physical or human processes, rare events, critical events such as failures and attacks.
  - Complexity of decision reflected in the complexity of the agent's decision process (goal management and planning) and impacted by factors such as diversity of goals and size of the space of solutions for planning.

- Additionally, building autonomous agents involves difficult systems engineering problems that are not related to the fact that the agents are intelligent.
  - Agents should be
    - integrated in complex cyber physical environments systems e.g. electromechanical systems
    - be able to harmoniously collaborate with human operators – It's not just an HMI problem!
  - In an autonomous system, the agents must be properly coordinated to achieve:
    - Symbiosis: the coordination of agents does not impede the achievement of their individual goals
    - Synergy: agents collaborate to achieve global system goals by demonstrating collective intelligence.

Autonomous Systems – From Agent Intelligence to Collective Intelligence



Autonomous Systems – Hybrid Autopilot Design

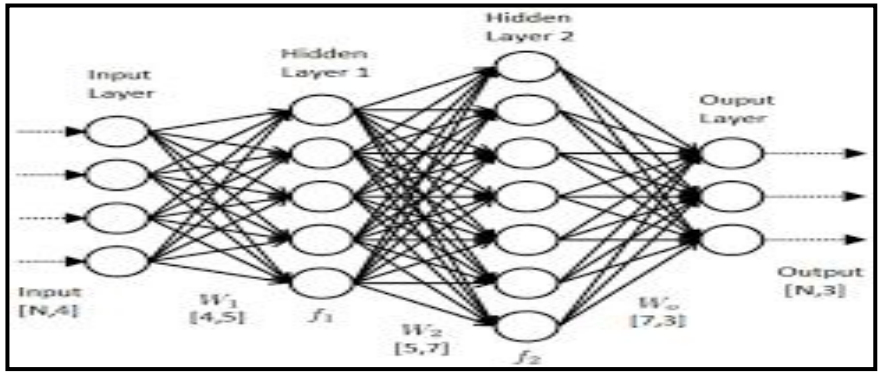


Overview

- Autonomous Systems
- Validation of Intelligent Systems
- Where Are We Going?

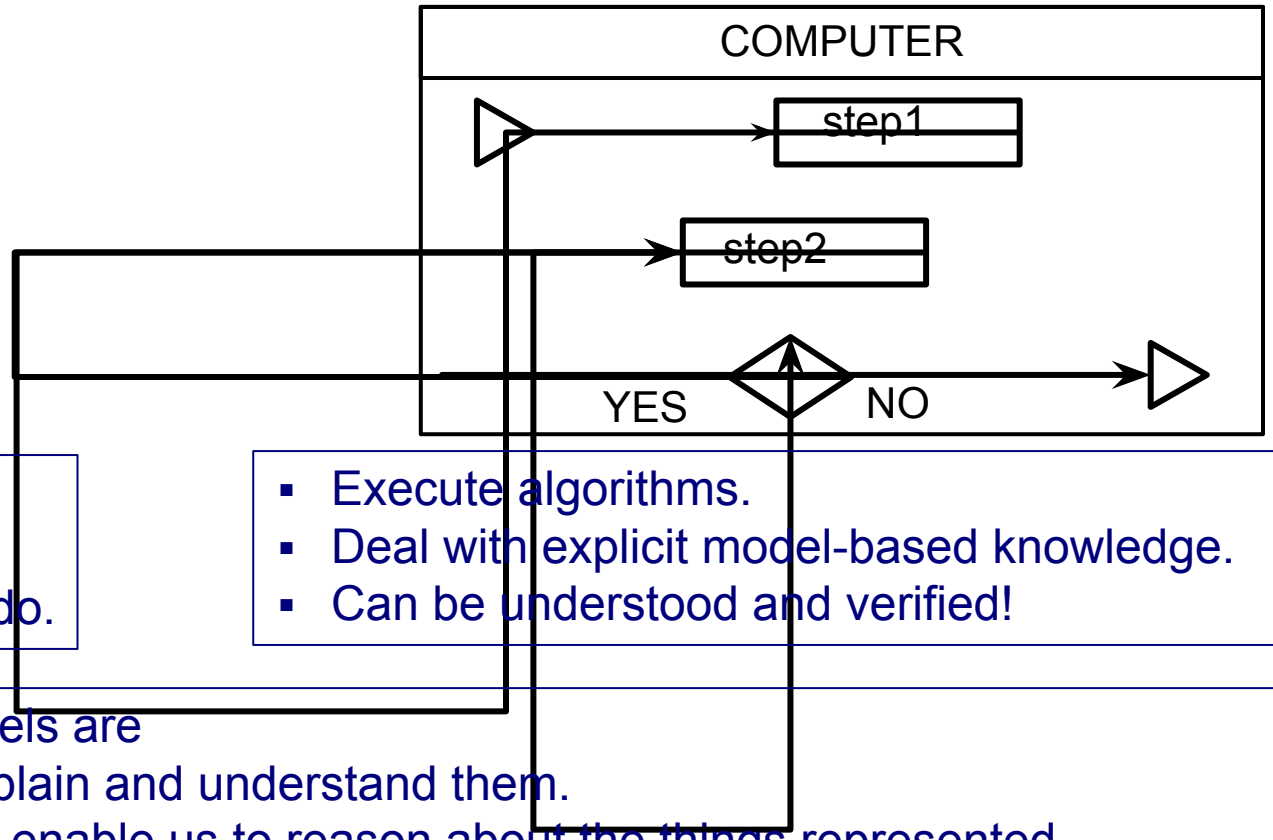
Validation of Intelligent Systems – Neural Networks vs. Traditional Digital Systems

NEURAL NETWORK



- Can be trained to generate data-driven knowledge
- Learn to separate "cats from dogs" as children do.

COMPUTER



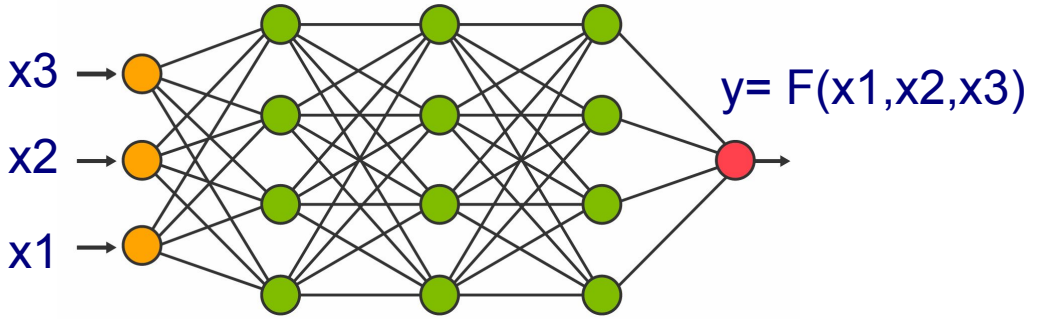
- Execute algorithms.
- Deal with explicit model-based knowledge.
- Can be understood and verified!

- Neural networks are artifacts, not models! Models are
  - representations of things that we use to explain and understand them.
  - essential for science and engineering: they enable us to reason about the things represented.
- Neural Networks do not execute algorithms, we use algorithms to train them!
- There is a remarkable analogy between the two computing paradigms and Kahneman's two systems of thinking:
  - System 1: fast automated thinking, dealing with implicit knowledge;
  - System 2: slow conscious thinking, dealing with explicit knowledge.

Validation of Intelligent Systems – Explainability

- A system is explainable if its behavior can be described by a model that lends itself to reasoning and analysis. System models are usually built following a compositionality principle:
  - In scientific disciplines, explainability is based on mathematical models, such as differential equations and statistical models.
  - For traditional digital systems, explainability is usually based on discrete models, such as transition systems.

- NN explainability : characterize the I/O behavior of a NN by a model obtained as the composition of the behavior of its elements.



  - For feed-forward networks, it is theoretically possible to calculate the output as a function F of the inputs, given the functions calculated by each node: *φ(weighted_sum_of_inputs),* where φ is an activation function.
  - However, the approach does not scale up for NN's in real-life applications. Only for classes of small feed-forward NNs with simple activation functions, approximations of F can be computed.

Note: Other, weaker notions of explainability fail to provide rigorous characterization sufficient to guarantee safety

Validation of Intelligent Systems – About Testing

There are two approaches for system property validation:

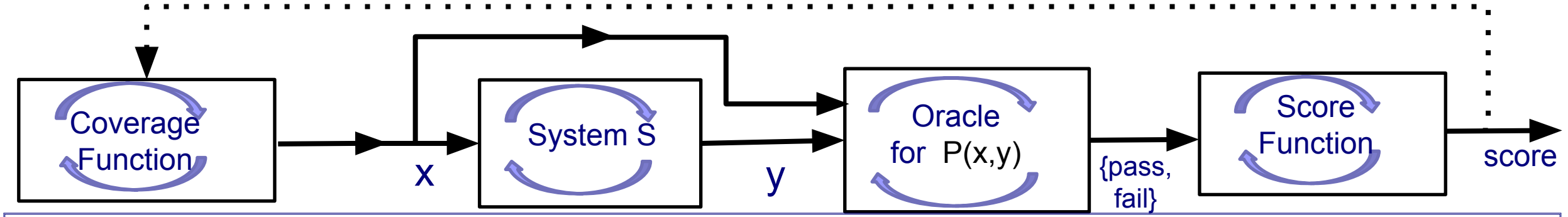
1) verification (by reasoning on a model);
2) testing (controlled experiment).

- Verification allows validation of system properties on a behavioral model of the system.
  - It is used to provide strong trustworthiness guarantees by reasoning on a global system model for properties such as safety, security, reliability.
  - It cannot be applied to neural networks in the current state of knowledge.

- Testing is an essential part of the scientific method applied to the production of empirical knowledge in all disciplines.
  - It is used to validate the observed behavior of a system in response to external stimuli; properties that cannot be captured as an I/O relationship cannot be tested,
  - It does not guarantee the validity of a safety property, only its non-falsification.
  - It is the basic empirical validation technique for traditional digital systems, for which a variety of test methods are available, such as structural, functional and metamorphic testing.

Validation of Intelligent Systems – Test Methods

- Tests are used to validate experimentally that a system y=S(x) satisfies a property P(x,y).
  1. System S: the system under test e.g. an electric bulb, an autopilot or an AI component;
  2. Property P: a predicate (hypothesis) characterizing the I/O behavior of S;
  3. Oracle: is an agent that can decide logically or empirically whether P(x,y) holds producing verdicts *pass* or *fail.*



- Test method: How do you choose between possible test cases and decide whether the process is successful or not?

  1. Coverage Function: *coverage*(X)∈[0,1] measures the extent to which the set of test cases X explores the characteristics of the system's behavior in relation to the property P
  2. Score Function: *score*(X,Y) measures for a test set (X,Y) the likelihood that S meets P .

Reproducibility: If (X1,Y1), (X2,Y2) are two sets of tests then:

coverage(X1)=*coverage*(X2) implies *score*(X1,Y1) ~ *score*(X2,Y2)

Validation of Intelligent Systems – Applicability of Test Methods

System S	Property P (Hypothesis)	Test method	Oracle for P	Results **Evidence** that S satisfies P / **Reproducibility** of results
Solar System	Newton's Theory (Mathematical model for S)	Model-based coverage criteria	Measurements to check Newton's laws	Conclusive evidence/ Objectivity
Flight Controller	Safety properties (Mathematical model for S)	Model-based coverage criteria	Automated analysis of system runs	Conclusive evidence/ Objectivity
Population	Response to a medical treatment e.g. vaccine	Statistics-based clinical tests and setting	Expert analysis of clinical data	Statistical evidence/ Statistical reproducibility
Image classifier	Relation □ ⊆ IMAGES×{cat,dog}	Test method for IMAGES?	Human oracle/ justifiable criteria.	Statistical evidence? / Statistical reproducibility?
Simulated Self-driving systems	Formally specified properties e.g. Traffic rules	Test method for driving scenarios?	Runtime verification of runs for given scenarios	Statistical evidence? / Statistical reproducibility ?
ChatGPT	Q/A relations in natural language	Test method for natural languages?	Human oracle Subjective criteria	No objective evidence

- The application of test methods to intelligent systems
  - is limited to technical properties that can
    - be rigorously specified, which excludes Q/A relations for natural language transformers;
    - be observed, which excludes "human-centric" properties e.g., intentionality, belief, awareness.
  - is hampered by adversarial examples -- observationally equivalent test cases give different scores;

Validation of Intelligent Systems – When an Autonomous Vehicle is Safe Enough?

Waymo has now driven 10 billion autonomous miles in simulation

Darrell Etherington @etherington / 11:17 pm CEST • July 10, 2019

Comment

- The inability to build global system models limits system validation to simulation and testing.
  - Simple simulation is not enough - how a simulated mile is related to a "real mile" ?
  - We need evidence, based on coverage criteria, that the simulation deals fairly with the many different situations, e.g., different road types, traffic conditions, weather conditions, etc.
- Test methods to calculate, on the basis of statistical analysis, confidence levels for given properties.
  - Sampling theory: methods for building sample scenarios that adequately cover real-life situations
  - Repeatability: for two samples of scenarios with the same degree of coverage, the estimated confidence levels are approximately the same

Overview

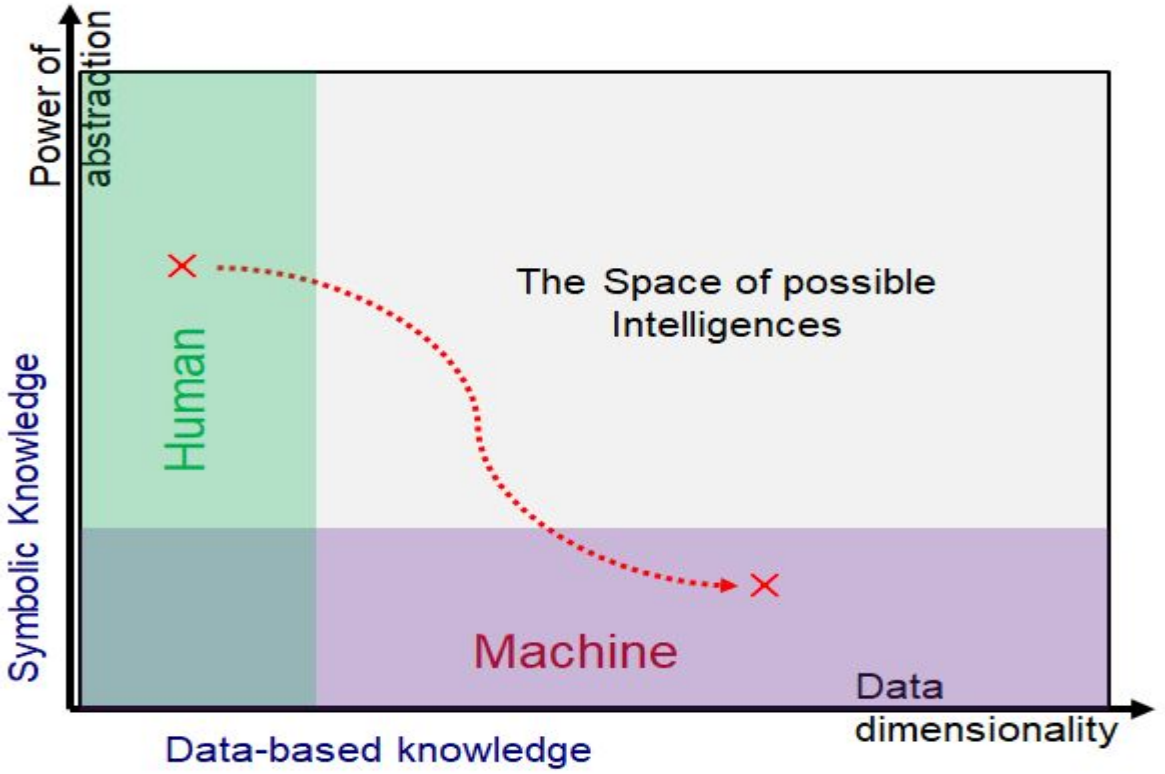
- Autonomous Systems
- Validation of Intelligent Systems
- Where Are We Going?

Where Are We Going? – The Space of Possible Intelligences

- Autonomous systems encompass a multi-faceted concept of intelligence.
  - There are multiple intelligences, each characterizing the ability to perform tasks in different contexts;
    To say that "S1 is smarter than S2" is meaningless without specifying the task(s) and the criteria for success.
  - Human intelligence is not "general purpose"; it is the result of historical evolution in a given physical environment.
    If human intelligence is the benchmark, AI should be able to perform/coordinate a set of tasks characterizing human skills.

- The space of possible intelligences: equivalent systems may use very different creative processes.
  - Humans are limited in analysis of multidimensional data, but are capable of common sense, abstraction and creativity.
  - AI systems outperform humans in learning multidimensional data, but fail to link symbolic to data-based knowledge.

- We need to explore the vast space of intelligences, particularly by delving into the various aspects of human symbolic intelligence and their relationship to data-driven intelligence.
  - Can we bridge the gap between symbolic and concrete knowledge exclusively by using neural networks?
  - Is it possible to trade symbolic reasoning capability for data-based learning as shown by LLM's opening the way to efficient solutions to symbolic reasoning problems e.g. MathPrompter



Where Are We Going? – Property Validation of AI Systems

- Tendency to ignore the limits of intelligent system validation resulting from established systems engineering criteria.
  - Many works on “Ethical AI” superficially attribute mental attitudes such as belief, desire and intention to autonomous systems: “*we cannot show that an agent always does the right thing, but only that its actions are taken for the right reasons*”.
  - However, testing ethical properties you need at least that
    - the system is to some extent explainable to model conflict resolution between alternatives of a dilemma;
    - ethical properties can be monitored and tested on the system’s I/O behavior, that is the system is not black box.

- When it is impossible to apply the scientific method, we have to study specific techniques between rigorous validation and qualification tests for assessing human skills.

- What if we applied qualification exams rather than rigorous tests to LLMs?
  After all, there is every reason to believe that LLMs will be able to pass the final exams just as well as students.
  However, we must not ignore two fundamental differences between NNs and humans:
  - Human thinking is robust, whereas neural networks are not (slight changes in questions imply different answers).
  - Human thinking based on common-sense knowledge, is better placed to avoid inconsistencies in the answers produced.

- ✔ Avoid religious debates about the allegedly human-centric properties of machines, without having explored the extent to which their rigorous validation is possible.

- ✔ Strive to overcome current limitations with clarity, developing new foundations, and possibly revising epistemic and methodological requirements, where necessary.

Where Are We Going? – Technology and Anthropogenic Risks

- AI is not a threat, threats come from our inability to use AI wisely.
  - Like any technology, AI opens the way to new achievements. Atomic energy can be used to generate electricity, while nuclear weapons can destroy humanity.
  - Talking about the inevitable domination of AI over man obscures the debate about our responsibility in the use of these technologies.

- Technology risk: associated with hazards compromising the system's ability to meet technical requirements, in particular safety or security risks
  - Safety means that the system can withstand any type of incident that could place it in "critical situations" that could harm the human and physical environment.
  - Security means that the system is resilient to any malevolent action by unauthorized users that could deliberately misuse the system in particular threatening data integrity, privacy and system availability.

- For traditional systems, safety and security are properties implied by technical requirements driving system development.

  The transposition of these concepts into AI systems seems problematic:
  - If LLM safety consists in not providing knowledge of a certain type, it is difficult to achieve this by design.

Anthropogenic risk: arising from intentional or unintentional misuse of technology violating regulatory or legal frameworks.
- Even if an LLM can generate deepfakes – that may be considered as violation of a safety requirement – using LLMs to produce deepfakes can be prohibited by law!

Where Are We Going? – Managing Technology and Anthropogenic Risks

- Technology Risks: The trustworthiness of all kinds of artefacts, from toasters, to toys, buildings, planes and cars.
  - is determined by standards relying on scientific and technical knowledge;
  - is controlled by independent bodies overseen by government agencies, e.g., in the US, FDA, FAA, NHTSA.

- Unfortunately, ICT systems and applications are not subject to this general rule requiring security and safety guarantees.
  - Exceptions are some critical applications (transport, nuclear power plants...).
  - Today, for AI applications, the lack of standards is compounded by permissive policies e.g. competent US authorities accept, for autonomous cars and medical devices, "self-certification" by manufacturers!

- Anthropogenic Risks: We need a framework that prohibits offences against human dignity e.g. impersonation, deepfakes, and regulates the use of AI weighting the involved risks against the benefits obtained.

  - Protection of privacy vs. security

  - Freedom of choice vs. performance: not to give decision-making power to systems if we are not sure that
    - they use reliable information in an unbiased and neutral way;
    - the gain in performance is commensurable with the lack of human control.

  - Division of work between human and machine: technological progress and innovation imply a loss of skills
    - The use of levers has made muscle power less necessary for our survival.
    - The lack of muscle power is not as dramatic as the loss of the skill to produce knowledge and act responsibly - which is the essence of human nature.

    Even if computers cannot surpass human intelligence, it is possible that humans could become too dependent on machines, overwhelmed by the complexity of managing them, or by laziness and convenience.
    And that would be a bad scenario for mankind!

Where Are We Going? – AI meets Systems Engineering

- The development of autonomous systems requires a marriage between ICT and AI, which poses non-trivial technical problems. New trends are disrupting traditional critical systems engineering.
  - adopting ML-based end-to-end solutions that do not provide trustworthiness guarantees;
  - allowing "self-certification“, in the absence of standards;
  - allowing regular updates of critical software - trustworthiness cannot be guaranteed at design time as required by standards - systems will be evolvable, with no end point in their evolution.

- Hybrid design leveraging on a solid body of knowledge for safe and efficient decision making.
  - Getting around the non-explainability obstacle: Build trusted systems from untrusted components.
  - Linking symbolic and non-symbolic knowledge e.g. sensory information and models used for decision-making.
  - For AI systems
    - Consider how restrictions on training data sets allow for better predictability and controllability:

      when an LLM explains how to make a bomb, it sums up information acquired during its training
    - Explore new avenues for explainable AI.

- System validation marked by the shift from rationalism to empiricism.
  - Simple simulation is not enough - Develop statistical testing techniques for AI monitors and end-to-end controllers.
  - Weaker trustworthiness guarantees that can be offset by the use of knowledge-based techniques.

- The transition from Automation to Autonomy cannot be progressive! We need to develop a new scientific and engineering foundation. And this will take some time.

Where Are We Going? – Exploring Human Consciousness

- Our progress in building intelligent systems will ultimately depend on our ability to elicit the mind-brain relationship.
  - Mental phenomena are essential for understanding human behavior, just as software is very important for understanding what a computer does!
  - Many will claim that the mind-brain problem can only be the subject of philosophical inquiry, and that a strict "scientific" approach is impossible or even misplaced.
  - Human intelligence cannot be unravelled if we focus exclusively on the study and simulation of brain processes and ignore mental phenomena, see the Human Brain Project.

- Questions about cosmogony or the evolution of species seem easier in the face of the "big bang of consciousness":
  - *How did cognition and language appear through evolution? How do we comprehend by linking phenomena to analogies and metaphors? How do we ascribe meaning to symbols? How do we create? How does awareness function as a system that combines wishes, motives and will? How do we choose goals and act on them? …*

- Linking mental functions and brain networks is a worthwhile, imperative endeavor where interdisciplinary cooperation between informatics, biology and medicine, could be decisive.

- Let's hope that one day we will be able to unravel the mystery of the "big bang of consciousness", and that physical cosmology will be complemented by a parallel one, which will shed light on our unexplored self and make us climb a little higher on the marvellous scale of self-knowledge.

Thank you