# AI: Where We Are, Where Are We Going?

Joseph Sifakis

Verimag Laboratory,
Grenoble, France

7th **A**dvanced **C**ourse on **D**ata Science & Machine **L**earning
Riva del Sole Resort, June 10, 2024

# AI – Where We Are, Where Are We Going?

At present, there's a great deal of confusion as to the final objective, fuelled by the media and large technology companies, who, through grandiose large-scale projects, spread opinions suggesting that human-level AI is only a matter of years away.

Opinions divided between two very different positions:

- Some AI research and companies such as OpenAI and DeepMind see Artificial General Intelligence (AGI), an ill-defined term, as the ultimate goal
  - suggesting that it can be achieved through machine learning and its further developments – ML is the “end of the story” whatever the result is.
  - 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.

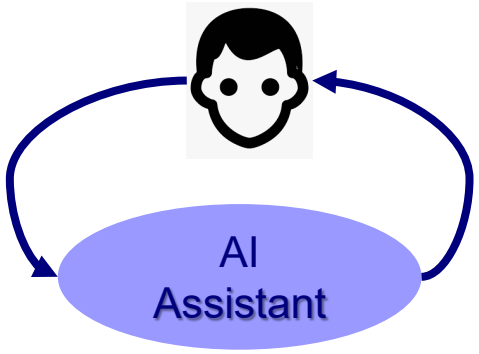
- Others see the goal of AI as building machines with Human-level Intelligence, which requires agreement on what human intelligence is and, more importantly, on methods for comparing human and machine intelligence.
  - According to the Oxford dictionary, intelligence is defined as
    *“the ability to 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.

Without a clear idea of what intelligence is, we cannot develop a theory of how it works!

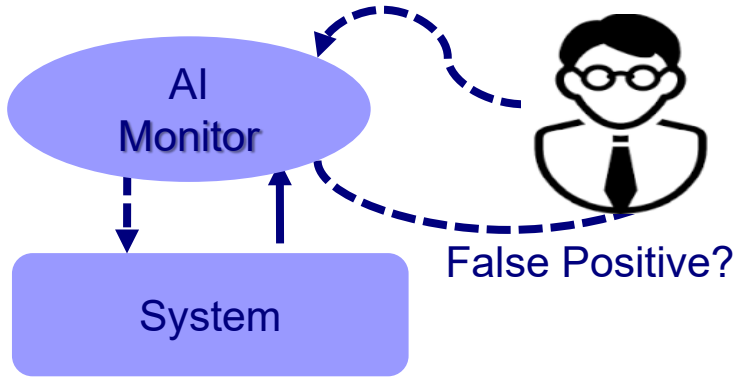
# Where We Are? – From Conversational to Autonomous AI

- AI is still in its infancy, despite impressive results culminating in the arrival of generative AI,
  - it only gives us the elements to build intelligent systems, but we don't have the principles and techniques to synthesise them, for example in the way we construct bridges and buildings..
  - It focuses on assistants, while its future applications require continuous interaction with little or no human intervention.

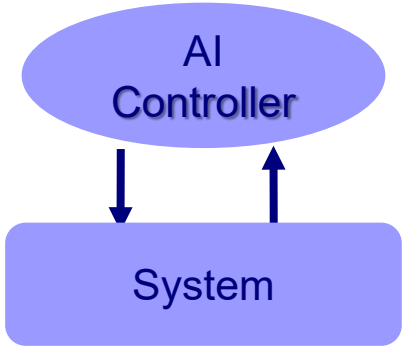
- Three different ways to use AI systems:
  1. Assistants that in interaction with a user, provide a given service;
  2. Monitors of a system behavior synthesizing knowledge to detect or predict critical situations;
  3. Controllers of a system so that its behavior meets a given set of requirements,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




Conversational AI




Predictive/Analytical AI




Autonomous AI

- The AI industry revolution has only just begun! Its realisation depends largely on our ability to develop AI monitors and end-to-end AI controllers to build autonomous systems.

# Where We Are? – Validation 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.
  - In particular, Safe AI has been the subject of international summits, deliberations by the UN - Specialised research institutes have been set up in the United States and the United Kingdom, with Korea and France soon to follow.
  - Safety is just a concept to be instantiated according to the type of AI and its application: LLM safety $\neq$ Autopilot safety.

- In addition to technical properties, 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 validation methodologies to AI systems? If so, what are the obstacles?

# Overview

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

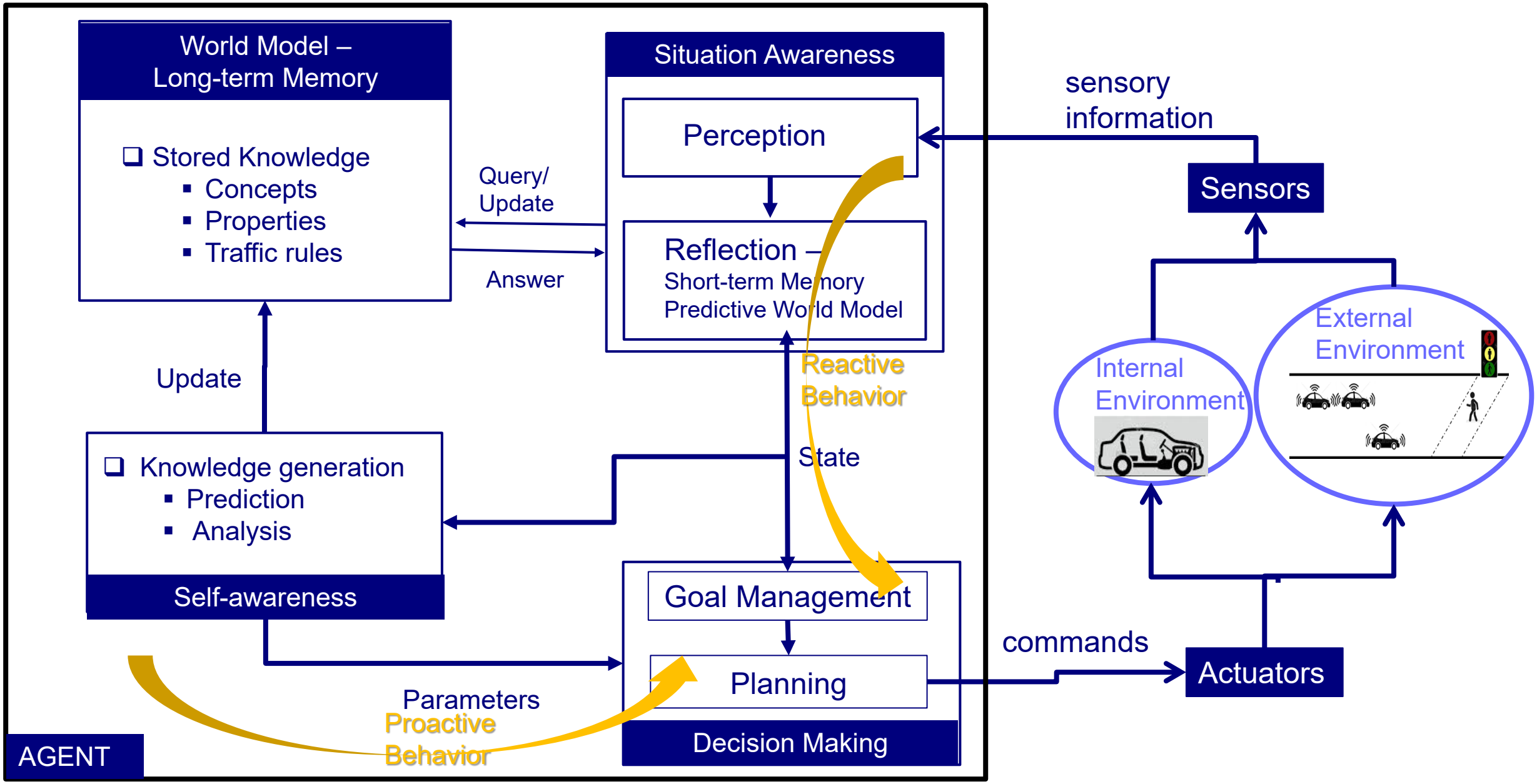
# Autonomous Systems – Trends and Characteristics

- Autonomous systems are the ultimate stage in the evolution of AI,
  - are distributed real-time systems, made up of agents, each pursuing specific goals (individual intelligence), but all coordinating in such a way that the system's behavior meets specific goals (collective intelligence).
  - are often critical systems intended to replace human operators in complex systems and organizations as envisioned by the IIoT, e.g. self-driving cars, smart grids, smart factories, robotic process automation.
  - are highly dynamic, reconfigurable systems that never stop and evolve online to adapt to the constantly changing environments and user requirements – design-time vs. runtime correctness.

- Current industrial trends in the development of autonomous systems are reinforced by the advent of generative AI, but obtaining trustworthiness guarantees remains an unavoidable and challenging objective.
  - The autonomous car sector is leading the way with AI-based end-to-end solutions that lack trustworthiness guarantees;
  - Progress in other sectors such as avionics, robotics, and networks, is more gradual.

- To make the autonomy vision a reality, new theoretical and technical foundations need to be developed,
  - that rejuvenate traditional systems engineering with hybrid design flows supporting the integration of data-based and model based techniques seeking trade-offs between design-time and run time correctness.
  - where knowledge development and management is an additional and decisive factor in overcoming complexity and compensating for the non-explicability of AI.

# Autonomous Systems – Self-driving Agent Architecture



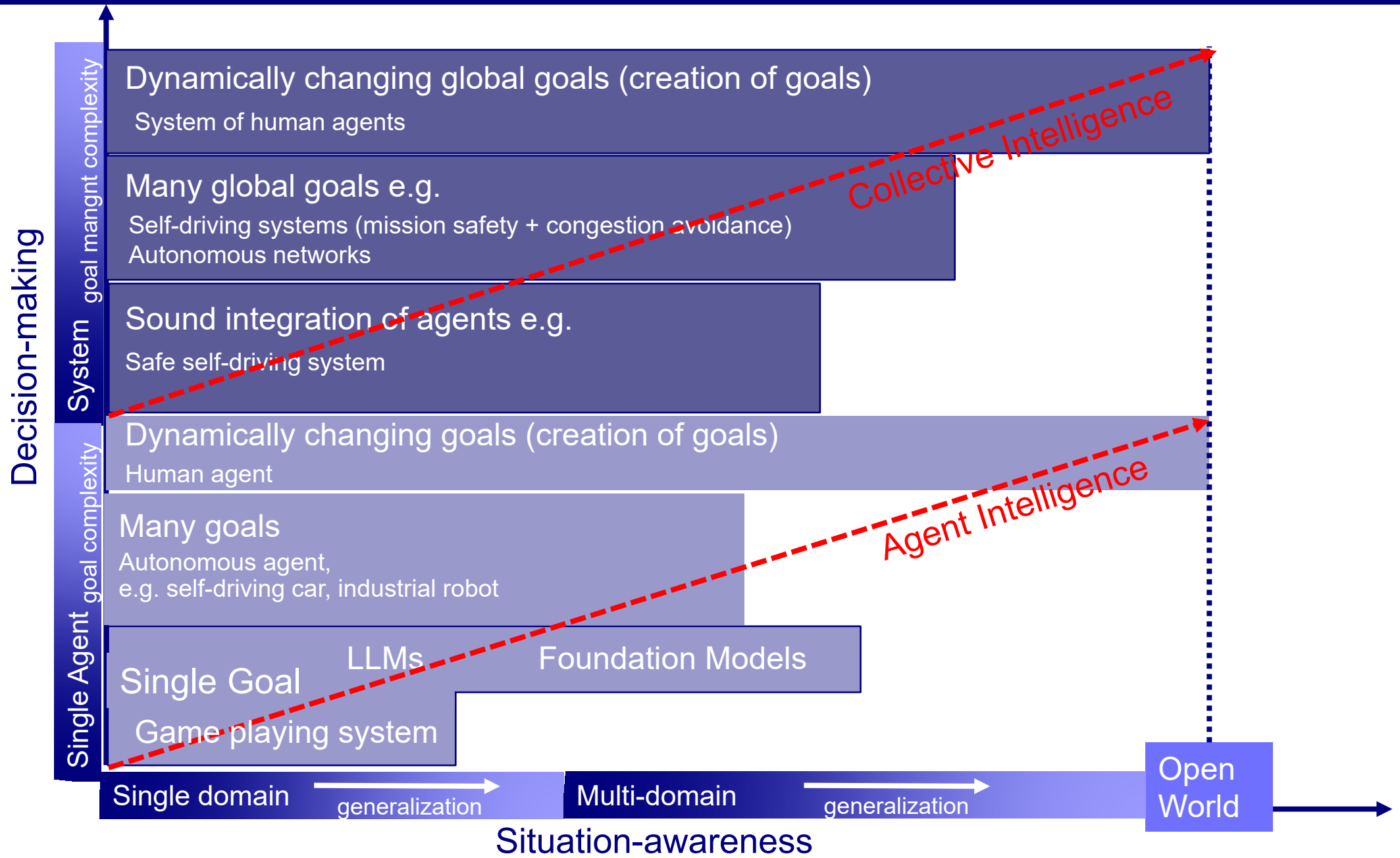
# Autonomous Systems – Complexity Issues

- Autonomous agents rely on computational intelligence to overcome complexity limitations
  - 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.

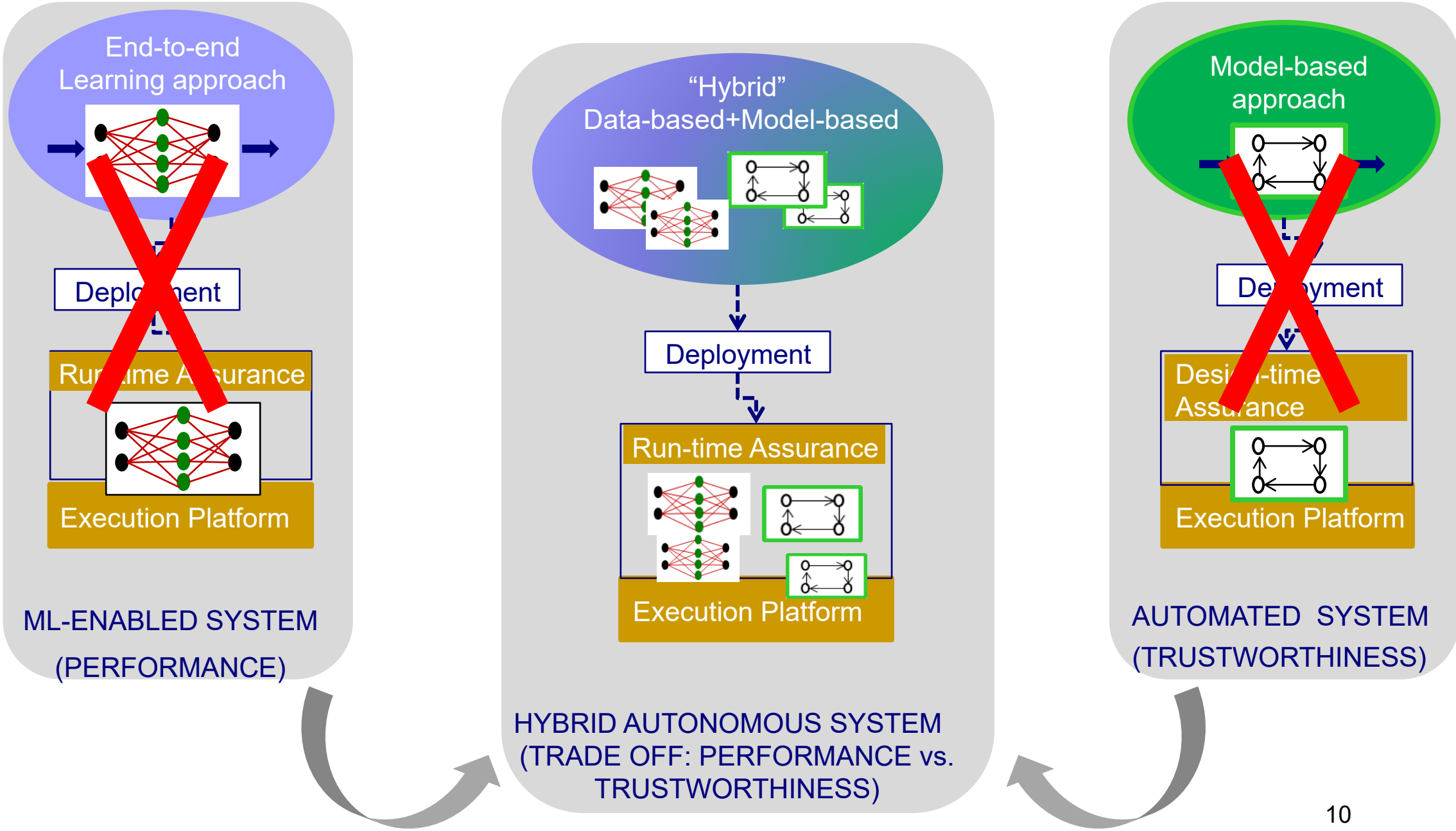
- However, building autonomous agents is not enough!
  - Agents should be
    - integrated in complex cyber physical environments systems e.g. electromechanical systems
    - be able to harmoniously collaborate with human operators – It's not just an HMI problem!
  - Agents of a system should be adequately 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.

See the controversies surrounding the deployment of robotaxis in SF, e.g. obstructing traffic, blocking police cars.

# Autonomous Systems – From Agent Intelligence to Collective Intelligence



# Autonomous Systems – Agent Design

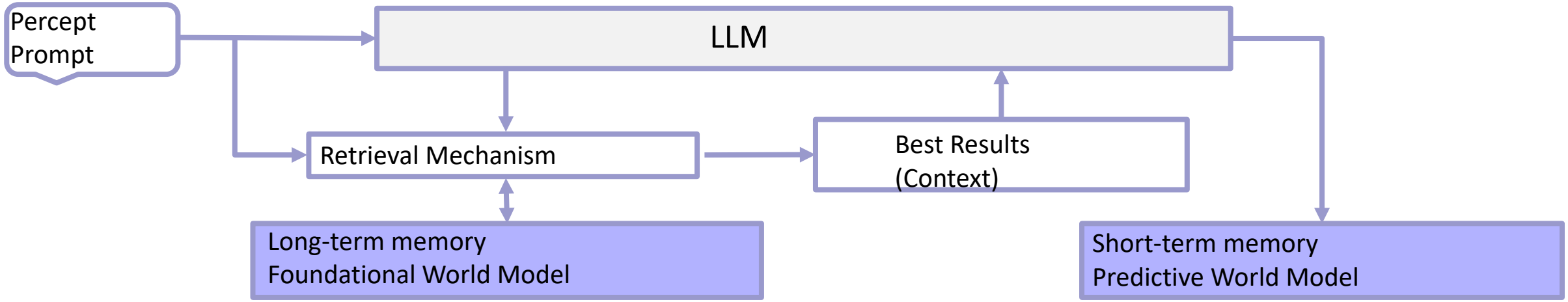


# Autonomous Systems – Agent AI

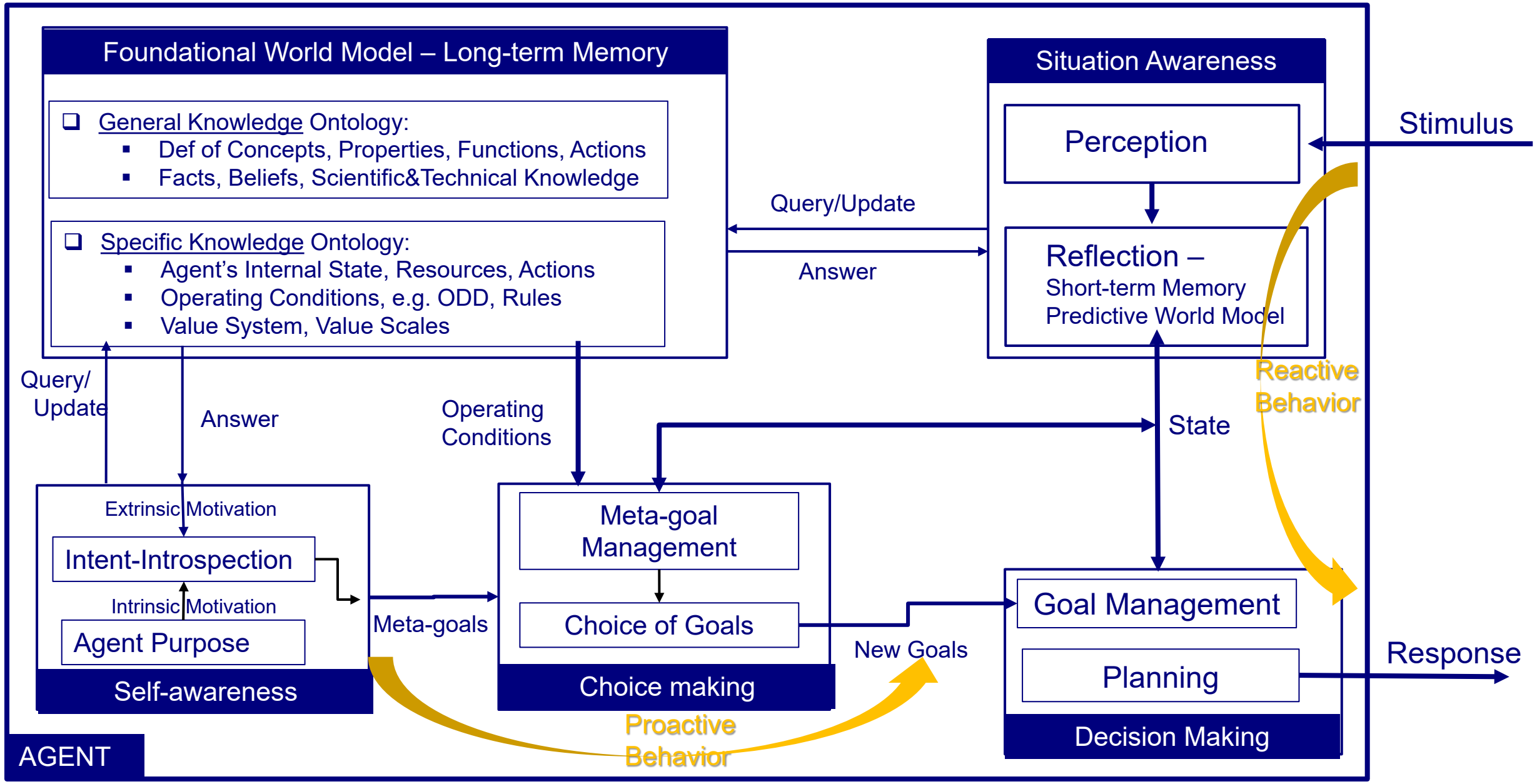
With the advent of LLMs, AI agents have attracted growing interest.

- **Move from monolithic E2E solutions to architectures with World Models**
  - to estimate missing information about the state of the world not provided by perception;
  - to predict plausible future states of the world.
- **Linking ML and symbolic computation** is essential to achieve autonomous AI
  - Approach 1: symbolic computation can emerge from learning with increasingly powerful machines ("scale is all you need!").
  - Approach 2: symbolic reasoning must be hard-coded from the outset, e.g. neurosymbolic techniques.
- **LLMs are not enough:** linking to domain specific knowledge for accuracy and predictive power.
  - LLMs grounded to symbolic engines such as AlphaGeometry, WolframAlpha, simulators, probabilistic programming tools…
  - LLMs use World Models stored in long-term memory, e.g. Retrieval-Augmented Generation (RAG)

**RAG Architecture integrating an LLM and World Models**



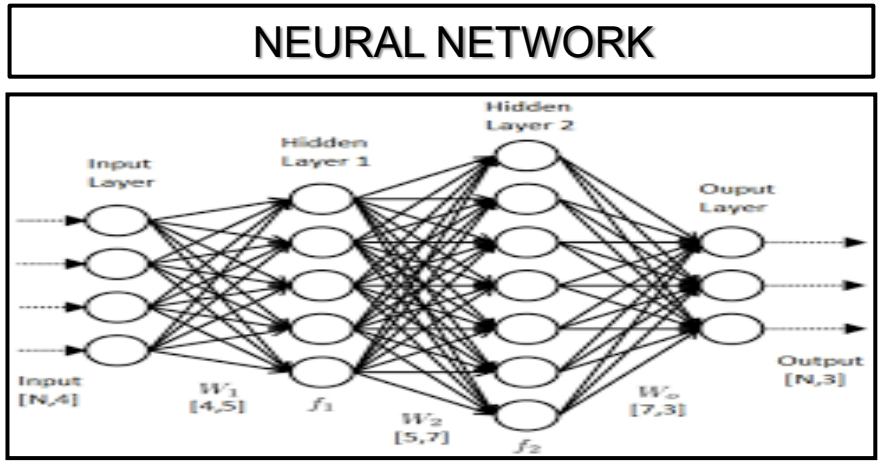
# Autonomous Systems – General Agent Architecture



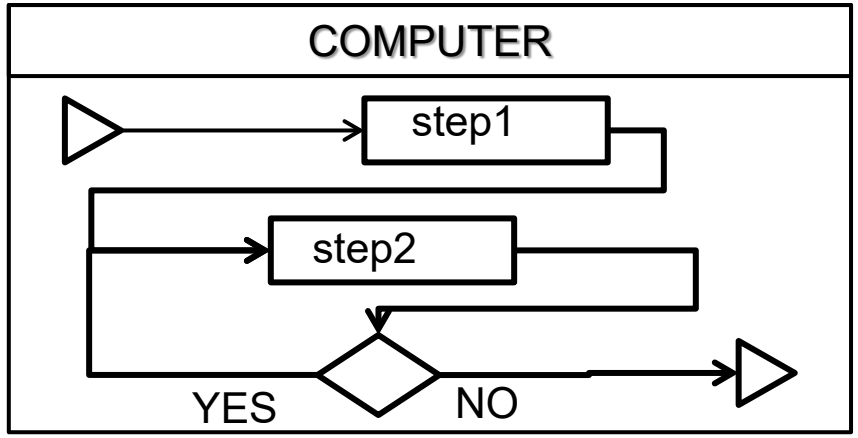
# Overview

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

# Validation of AI Systems – Neural Networks vs. Traditional Digital Systems




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



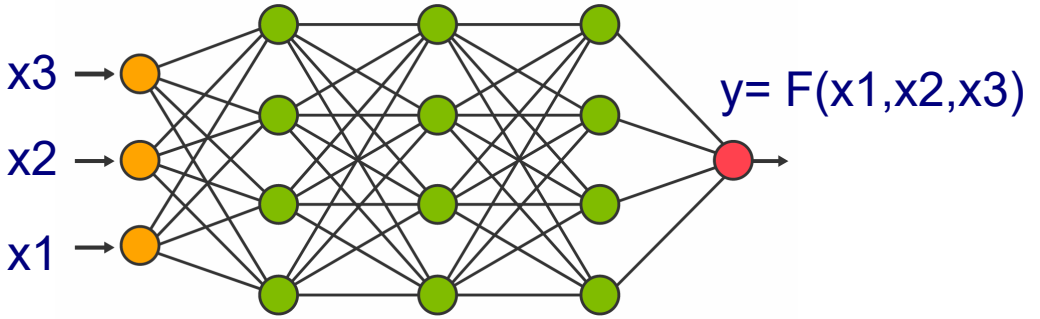

- Execute algorithms.
- Deal with explicit model-based knowledge.
- Can be understood and verified!

- 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 AI 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 interpretability fail to provide rigorous characterization sufficient to guarantee safety properties, e.g., extracting a textual description of behavior or causal dependency between input/output variables.

# Validation of AI Systems –Validation Methods

- Epistemic and methodological imperatives applied to the development of scientific and technical knowledge, establishing that "a system S satisfies a property P",
  - require that not only P be rigorously defined, but also that a falsifiable validation method be provided.
  - combine reasoning on models (verification) and experiment (testing).

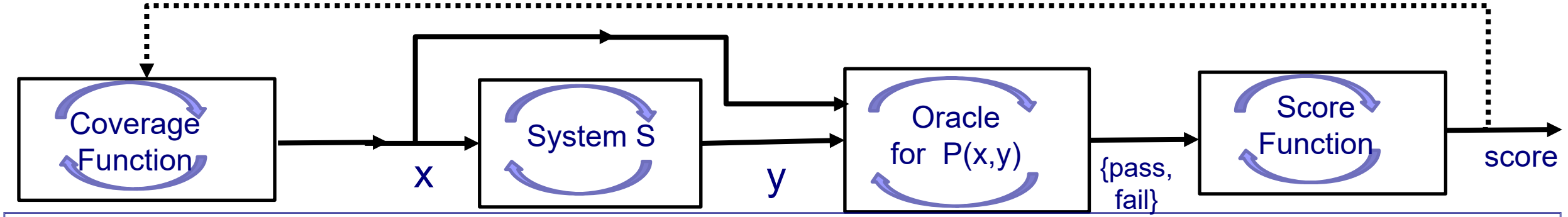
- Verification consists in comparing a model of S against some specification of P.
  - can examine the whole system behavior described by a model, and decide about the validity of its properties.
  - can validate properties involving universal quantification e.g. that all system states are safe, or that for any system there exists a rejuvenation state.
  - is the only way to obtain solid guarantees on the satisfaction of technical properties.

- Testing is a controlled experiment on the S (real or virtual) to assess the degree of validity of P.
  - is subject to observability and controllability constraints: distinction between system inputs (controllable) and outputs (observable)
  - is limited to properties P characterizing an I/O relation; properties involving universal quantification such as safety and security, can only be falsified.

To what extent can the properties of AI systems be validated using test methods?

# Validation of AI 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)$\in$[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 AI Systems – Applicability of Test Methods

| System S | Property P | Test method | Oracle for P | Results<br>**Evidence** that S satisfies P/<br>**Reproducibility** of results |
|---|---|---|---|---|
| Solar System | Newton's second law F=ma | Model-based coverage criteria | Method to check that F=ma | Conclusive evidence/<br>Objectivity |
| Software | y=S(x) $\forall x \in Dom(x)$ P(x,y)<br>P: correctness property | Model-based coverage criteria | Automated analysis for a given set of test cases | Conclusive evidence/<br>Objectivity |
| Population | Validating the effect of a medicine on the population. | Statistics-based clinical tests and setting | Expert analysis of a sample of clinical data | Statistical evidence/<br>Statistical reproducibility |
| Image classifier | Relation<br>$\rightarrow \subseteq$ IMAGES×{cat,dog} | Test method for IMAGES? | Human oracle/ for an adequate sample coverage | Statistical evidence? /<br>Statistical reproducibility? |
| Self-driving system | P(x,y): safety property<br>$\forall$x scenario and corr. trajectory y | Test method for driving scenarios? | Runtime verification for an adequate sample coverage | Statistical evidence? /<br>Statistical reproducibility ? |
| Q&A system e.g. ChatGPT | Q/A relations in natural language | Test method for natural languages? | Human oracle<br>Subjective criteria | No objective evidence |

- The development of test methods for AI systems
  - is hampered by the lack of explainable models for reasoning about the test space and developing coverage criteria.
  - is limited to properties P that can
    - be rigorously specified, which excludes Q/A relations LLMs;
    - be observed, which excludes "human-centric" properties e.g., intentionality, belief, awareness.

# Validation of AI Systems – When a Self-driving Car  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.

- We are badly lacking test methods for AI systems similar to those applied to software and hardware systems.
  - Sampling theory: methods for building samples that adequately cover real-life situations.
  - Repeatability: for two samples with the same degree of coverage, the estimated confidence levels are approximately the same..

# Overview

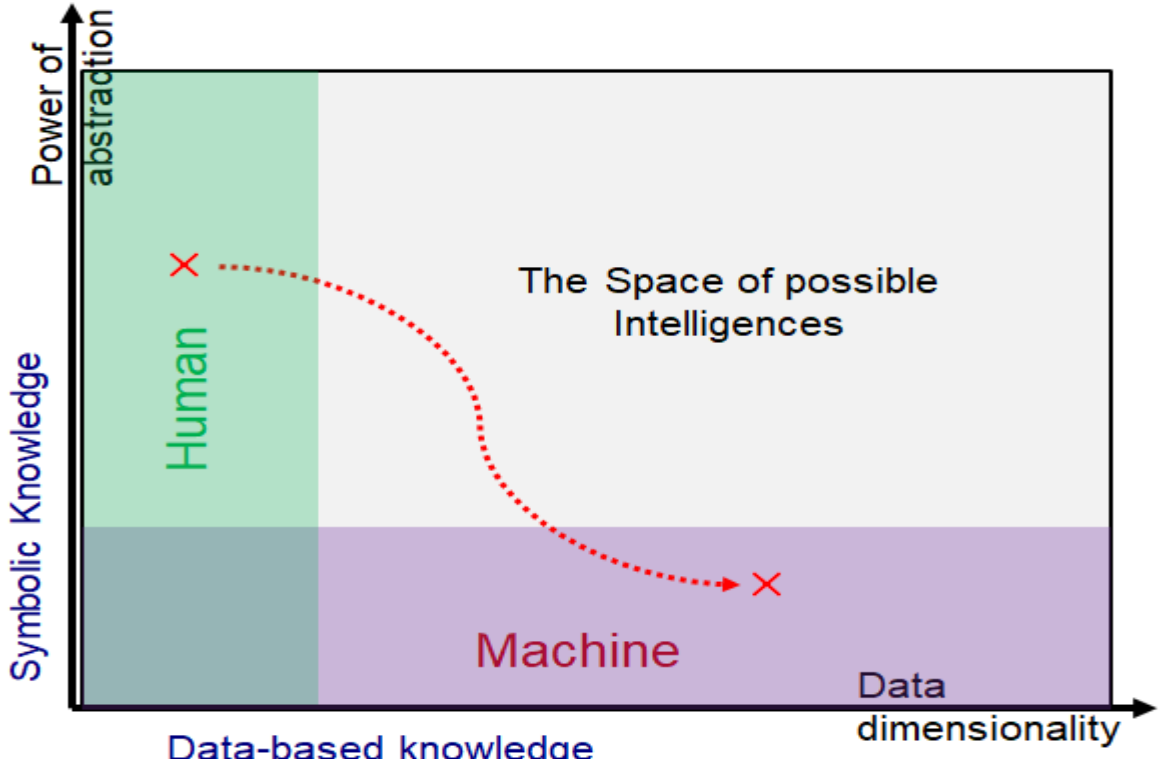
- Autonomous Systems
- Validation of AI Systems
- Where Are We Going?

# 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 a task in a given context;
    To say that "S1 is smarter than S2" is meaningless without specifying the task and the criteria for success.
  - Human intelligence is not a theoretical concept, 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

- The tendency to consider AIs as moral entities seems to be a way of avoiding the debate on the validity of technical properties and the application of associated scientific methods.
  - Unlike technical properties, ethical properties cannot be decided solely on the basis of the agent's observable behavior, without knowing its beliefs and intentions.
    - saying that “the earth is round” may be lie or ignorance;
    - non awareness that I am doing something wrong does not imply my responsibility.
  - Many works on “Ethical AI” superficially attribute mental attitudes such as belief, desire and intention to autonomous agents: “*we cannot show that an agent always does the right thing, but only that its actions are taken for the right reasons*”.

- 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 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.
  - Humans are responsible for the consequences of their actions or omissions - we won't put a machine in jail!!

- ✓ Avoid superficial debates about human-centric properties of machines, in the absence of any rigorous characterisation.

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

# 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 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 using hybrid architectures: 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!

- System validation marked by the shift from rationalism to empiricism.
  - Random testing 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? – Risks and Their Regulation

- Unfounded and deliberately nurtured myths predict domination of man by machine. However, AI is neither good nor bad! The challenge is to use it wisely, to prevent risks by regulating its application, and make the most of it for society.

  .

  - Technology risks: from hazards compromising an AI's ability to meet safety and security requirements.
  - Human risks: from misuse or unintentional impact of AI that can be controlled by regulatory or legal frameworks.
    - *Loss of jobs due to automation can be offset by an appropriate social policy e.g. for quality of life, new needs.*
    - *If an LLM can generate deepfakes - which is considered a technological risk - its use can be prohibited by law!*

- Official statements from governments and institutions affirm the need for AI regulation.

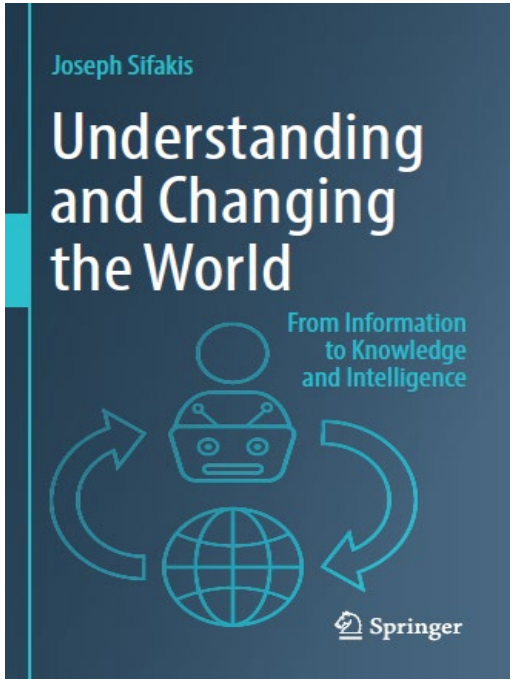
  However, there is no agreement on

  - What are the risks involved in using AI? What can and should be regulated? How can this be achieved in practice?

- There is a gap between the EU and the US when it comes to AI regulation:
  - EU has strong and more complete AI regulations, in particular with two texts, the AI Act and the Digital Services Act;.
  - US regulations are much less coercive e.g. the AI Executive Order (issued on October 30 2023).

The chances of reaching agreement on a global regulatory framework for AI, as advocated by the UN, are currently slim.

# Publications on Autonomous Systems




Joseph Sifakis. Understanding and Changing the World – From Information to Knowledge and Intelligence, Springer May 2022. https://link.springer.com/book/10.1007/978-981-19-1932-9


1. Joseph Sifakis. Autonomous Systems -- An Architectural Characterization, https://arxiv.org/abs/1811.10277, Nov 2018
2. Joseph Sifakis. System Design in the Era of IoT— Meeting the Autonomy Challenge, Methods and Tools for Rigorous System Design (MeTRiD 2018), EPTCS 272, 2018, pp. 1–22, doi:10.4204/EPTCS.272.1.
3. Joseph Sifakis. Can We Trust Autonomous Systems? Boundaries and Risks, ATVA, 2029, LNCS 11781, pp. 1–14, 2019. https://doi.org/10.1007/978-3-030-31784-3_4.
4. David Harel, Assaf Marron, Joseph Sifakis. Autonomics: In Search of a Foundation for Next Generation Autonomous Systems. PNAS, July 21, 2020, 117 (30) 17491-17498, https://doi.org/10.1073/pnas.2003162117
5. Joseph Sifakis David Harel. Trustworthy Autonomous System Development. ACM Transactions on Embedded Computing Systems, Volume 22,Issue 3, Article No.: 40, pp 1-24, April 2023, https://doi.org/10.1145/3545178
6. Joseph Sifakis. Testing System Intelligence. https://arxiv.org/abs/2305.11472, August 2023.

Thank you