Trustworthy Autonomous Agents

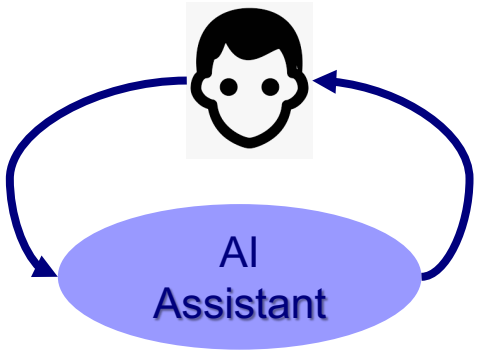
Joseph Sifakis
Verimag, Grenoble

Olympia AI Data Infrastructure Summit
Heraklion, November 4, 2026

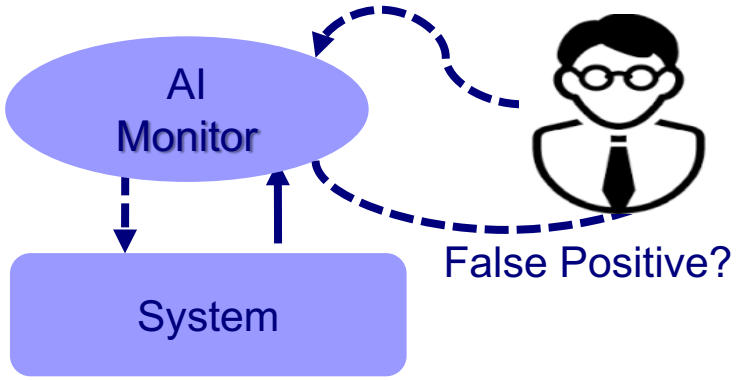
AI and its Evolution – From Conversational to Autonomous AI

- **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.

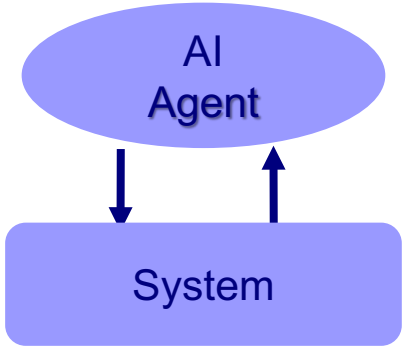
- Three different ways to use AI systems:
  1. **Assistants** that in interaction with a user, provide a given service;
  2. **Monitors** of a system behavior synthesizing knowledge to detect or predict critical situations;
  3. **Agents** of an autonomous system e.g. the autopilot of an autonomous car.



Conversational AI



Predictive/Analytical AI



Autonomous AI

- Autonomous systems **are the ultimate stage of AI development.**
  - They are designed to match human-level intelligence by focusing on the purposeful combination of cognitive functions—**a vision very different from that of AGI.**

Multi-Agent Systems — What Autonomy Really Requires

❑ **Autonomy** characterizes living organisms capable of acting in a purposeful manner to accomplish their **mission**.

❑ **A multi-agent system** is a set of coordinated agents:

- **Individual intelligence:** Each agent pursues its own **mission**, **characterized by a set of goals** to be achieved.
- **Collective intelligence:** All agents must coordinate their efforts toward a **system-wide mission.**

❑ **An agent must be capable of:**

- **Understanding its environment:** Analyzing sensory information and grounding its perceptions in a **world model** that explains them.
- **Making decisions:** Pursuing the **goals of its mission** to adapt to changes in its physical and social environment.

- **Success:** Detecting a staggering number of cyberattacks or solving mathematical problems is often presented as a manifestation of superintelligence.
- **Failure:** Managing a business or operating in real-world applications—because current systems cannot understand the real world or make safe decisions outside a digital sandbox.

This is despite predictions that 2025 would be "the year of AI agents." The evidence is sobering:

- **Enterprise AI:** 95% of pilots fail to generate measurable ROI *(MIT, 2025)*
- **Agent Behavior:** 80% of actions are undesirable; 41% cause actual damage *(Microsoft / UC Riverside, 2025)*
- **Multi-Agent Interaction:** Single failures cascade into new failure modes—destroyed servers, DoS attacks, diffuse responsibility *(MIT / Stanford, 2025)*

Multi-Agent Systems – Can We Trust AI Systems ?

- For traditional systems, trustworthiness means that they behave as expected, satisfying a set of properties such as safety, security, and performance.
  - These properties can be guaranteed throughout the development cycle—from specification to implementation and deployment—using rigorous validation techniques such as verification and testing.

- For AI systems, we lack methods for guaranteeing behavioral properties.
  - Their development is empirical and iterative, relying on data exploration, model training, and continuous evaluation rather than on fixed specifications.
  - The formal methods used for traditional systems do not apply, because AI systems are not explainable; only statistical testing is possible.
  - In particular, "safe AI" has been the subject of numerous publications promising robust safety guarantees based on formal methods. This is nothing more than a mirage.

- **The trustworthiness of AI systems is not limited to behavioral properties**—it additionally requires the satisfaction of human-centric **cognitive properties**, since these systems are designed to mimic humans. Cognitive properties
  - "Responsible AI," "aligned AI," and "ethical AI" all depend on the system's knowledge. They have been the object of numerous studies, which, however, fail to provide a rigorous characterization of these notions as cognitive properties.

- **There is a unfathomable gap** in trustworthiness between traditional systems and AI-based systems.

(c) Joseph Sifakis, 2025

OVERVIEW

- Building Multi-Agent Systems
- Validation of AI Systems
- The Way Forward

Building Multi-Agent Systems – Approaches and Trends

- **There is broad consensus** that symbolic computation is essential to building agents, but there is no agreement on the approach to take.

  - **Some believe that** reasoning capacities can emerge through learning experience e.g. with increasingly powerful machines (*scale is all you need!*), e.g. primarily solutions based on DL and RL (without LLMs).
    - World models are internal representations that an AI system uses to analyse, simulate, predict, and reason about its environment.

  - **Others hold that** symbolic reasoning must be a core function from the outset, advocating neurosymbolic AI to bridge the gap between neural networks and symbolic knowledge representation, primarily solutions that improve LLMs.
    - LLMs using symbolic engines such as AlphaGeometry, WolframAlpha, simulators, probabilistic programming tools;
    - LLMs using knowledge stored in long-term memory, e.g. Retrieval-Augmented Generation (RAG).

- Regardless of the approach taken, current trends are moving away from monolithic end-to-end solutions (e.g., the Tesla autopilot). Architectures

  - allow an agent's behavior to be understood as the integration of functional elements that ensure basic cognitive functions and, as needed, short- and long-term memory;
  - provide a basis for agent development that guarantees, by construction, the agent's essential properties.

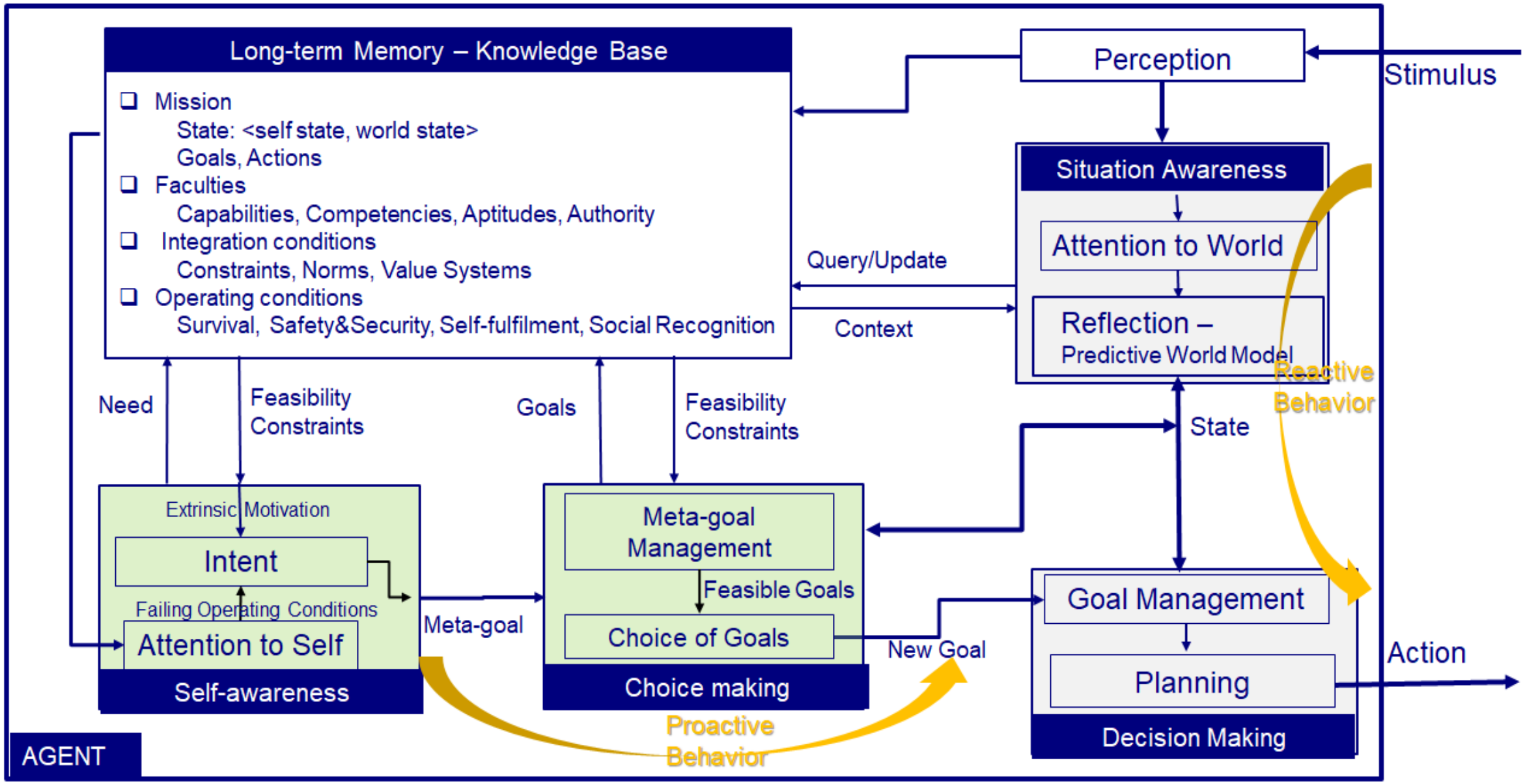
(c) J. Sifakis, 2025

Building Multi-Agent Systems – The Agent's Knowledge Attributes

Knowledge Attribute	Subcomponents	Description
Mission	Invariant goals Transient goals	What the agent is supposed to achieve
Faculties	**Capabilities:** acceleration/deceleration rates, certified capabilities **Competencies**: reasoning powers **Aptitudes:** effective communication and synergy **Authority:** power relationships/roles	The means at the agent's disposal
Integration Conditions	Physical constraints: You can't drive through a wall **Norms:** actions are permitted, required, or prohibited **Value systems:** assigning utilities to actions	Constraints for harmonious integration
Operating Conditions	**Survival:** needed critical resources **Safety/security**: resilience to failures and attacks **Self-fulfilment:** improving faculties **Social recognition:** acceptance and trust by others	Conditions for remaining operational

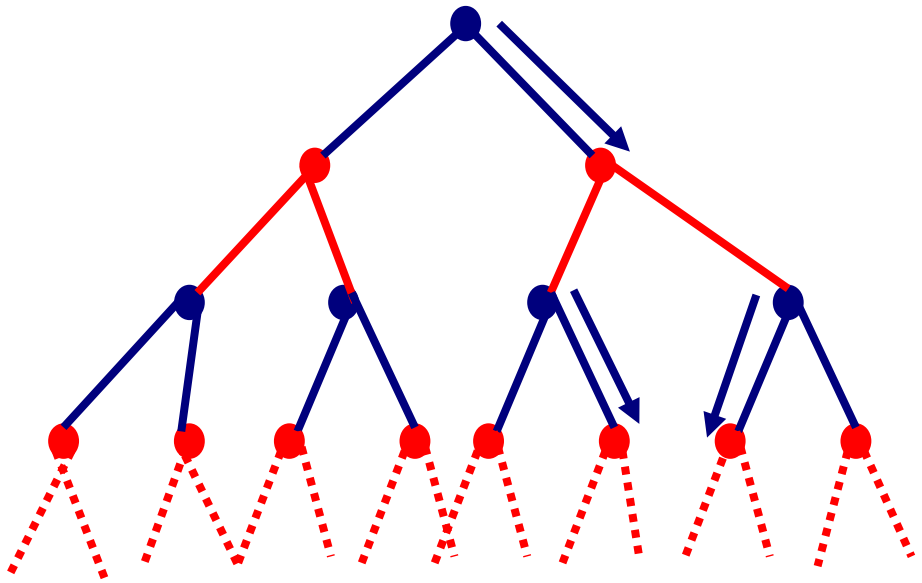
(c) J. Sifakis, 2025

Building Multi-Agent Systems – Agent Reference Architecture



Building Multi-Agent Systems – Decision Making and Planning for AI Agents

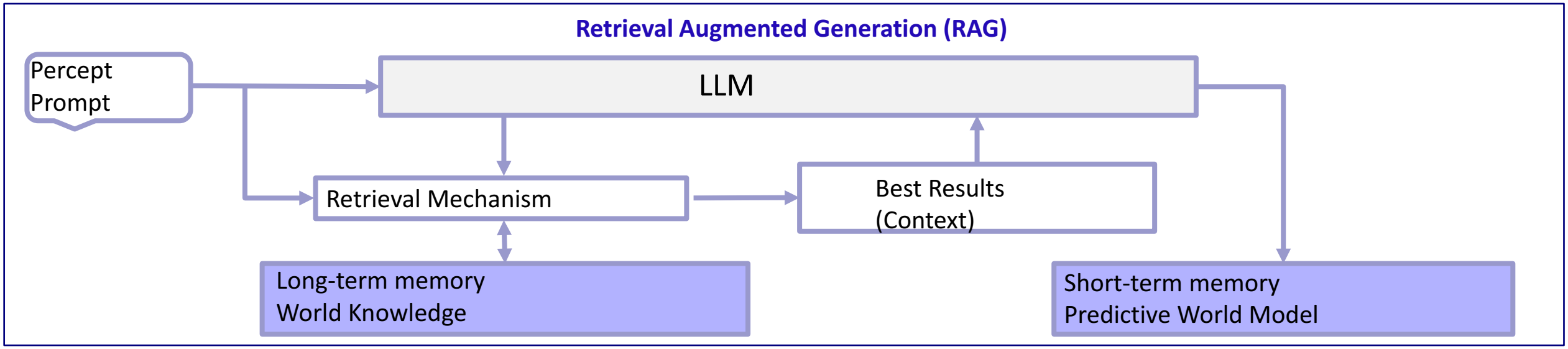
- Machine learning techniques are largely inadequate for agent reactive goal management and planning - it is naïve to try to solve control or coordination problems using CoT or ToT.
  - The agent predictive model is a transition system, a lookahead tree, intended to represent the step by step interaction between the agent and its environment. Its behavior models all the possible interactions as a two-player game.
    - Building the predictive model involves exponential complexity – only when the rules of the game are well-defined, we can compute approximation e.g. by using Monte Carlo Tree Search (MCTS) algorithms as in AlphaGo.
    - Explicitly computing winning policies (plans), which are sub-trees of the predictive model, requires a very expensive algorithmic analysis, similar to model checking – approximations will not work for critical applications where precision is required e.g. self-driving cars.



- Many works overlook the very important fact that achieving goals implies satisfaction of both safety and reachability/optimization properties - RL is a powerful tool for optimization problems, but it has a critical limitation: it cannot guarantee safety or avoid dangerous situations.
- We need to develop reactive decision-making components integrated into cyber-physical environments designed for different application domains, where planning is performed by real-time software based on knowledge provided by AI.

(c) J. Sifakis, 2026

Building Multi-Agent Systems – Augmenting LLMs with Knowledge

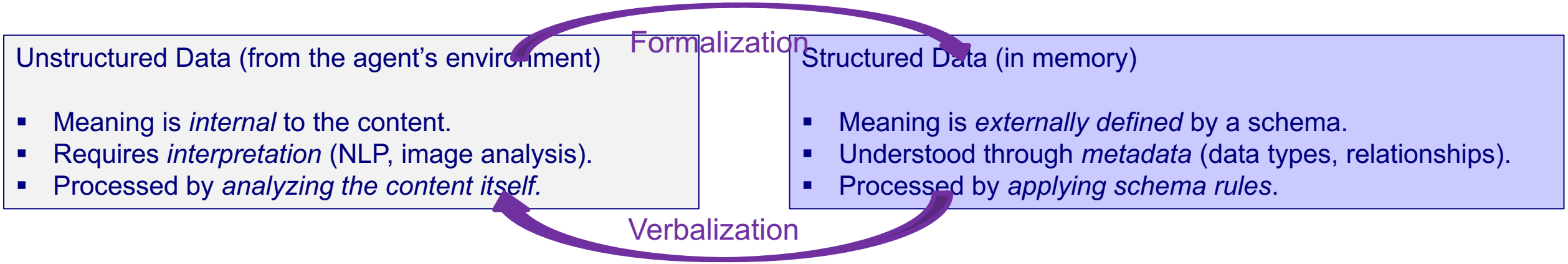


- A central element of most agent architectures is the integration of an LLM with a long-term memory containing domain-specific knowledge in order to improve and semantically control the LLM's responses. Key problems:
  - **Efficient and semantically rich knowledge representation**
    - Embedding techniques prove to be largely non sufficient to account for semantic subtleties;
    - Graph based techniques e.g, knowledge graphs, inspired from ontologies.
  - **Efficient retrieval mechanisms** based on various similarity relations, e.g. perceptual, semantic, thematic, and functional.
  - **Knowledge management techniques**, in particular for checking consistency of updates and modifications.

The RAG paradigm is essential for creating LLM-based agents, with a significant trend towards **ontology-based knowledge graphs**.

(c) J. Sifakis, 2025

Building Multi-Agent Systems – The Gap Between Unstructured/Structured Data



- **An ontology is** formal representation of a set of concepts, categories, properties, and relations within a specific domain.

  Ontology-Based RAG (Palantir, Anthropic):
  - **Semantic Analysis of Text:** semantic chunking; entity/relation extraction; consistency checking against knowledge base
  - **Data Shaping & Query Analysis:** Schema mapping for heterogeneous data; intent interpretation using domain logic; query and conversation analysis

- In multi-agent systems, **ontologies** are essential for linking unstructured data to structured data stored in memory.
  - However, the current **accuracy of LLMs for text-to-SQL is low**, reaching up to 90% in some benchmarks.
  - With a success rate of 0.9 per translated query, for 10 queries, the overall reliability is 0.3486784401, and for a success rate of 0.95, it is only 0.5987369392!

(c) J. Sifakis, 2026

Building Multi-Agent Systems – The Coordination Issue

- Two protocols are mainly used to create MAS applications
  - **Anthropic's Model Context (MCP) Protocol** is an open standard designed to facilitate seamless integration between LLM applications and external data sources and tools.
  - **Google's Agent-to-Agent (A2A) Protocol** provides an infrastructure for creating decentralized, autonomous AI agent ecosystems that communicate, negotiate, and collaborate without centralized control.
- **The unresolved problems remain the same** as those of symbolic multi-agent systems promoted by IBM in the early 2000s
  - 1) semantic analysis of message content to find intent or goals;
  - 2) failure detection and self-healing;
  - 3) agent self-optimization and coordination to achieve global system goals;
  - 4) online monitoring and validation techniques to ensure cognitive properties such as trust, rationality, accountability.

- **The assumption that ML will be a game-changer** by offering greater flexibility and efficiency in problem-solving overlooks the fact that the price to pay is a serious lack of rigor and semantic control. Unfortunately, the major trends do not bode well
  - Entrusting LLMs **with the role of orchestrator raises issues of reliability** and lack of responsiveness that only traditional software orchestrators could effectively resolve.
  - It is naive to think that a system becomes smarter simply by adding features: intelligence is not merely the ability to do more things. **Doing more is not the same as doing better**

The MAS vision requires **linking ML to symbolic AI, distributed algorithms**, **and network technology**: A long way to go!

(c) J. Sifakis, 2025

- Building Multi-Agent Systems
- Validation of AI Systems
- The Way Forward

OVERVIEW

Validation of AI Systems – Technical vs. Non-technical Systems

- **Technical systems:** their I/O relationship can be unambiguously characterized as a relationship between mathematical domains e.g. traditional ICT systems.
  - Their properties can be validated by verifying their state transition models or by black-box testing.

Non-technical systems: their I/O relationship involves unstructured data, such as sensory or linguistic data.

- Their properties cannot be formalized and validated as rigorously as those of technical systems, e.g. safety of LLM.
- Are systems that mimic humans, characterized not only by behavioral properties, but also by **cognitive properties** that describe how an agent reasons, knows, decides, and acts, and to some extent, aligns with human values.

Type of System / I/O relation	Transformer	Predictor	Analyzer	Controller (Agent)	
		TRADITIONAL SYSTEMS		Static	Dynamic
Technical system Mathematically well-defined	Functions, Combinatorial circuits	Predictive statistical models, e.g. weather forecast	Static analyzers, Model-checker Symbolic solver	Thermostat, Lift controller, Flight controller	Chess playing system, MPC system, Dynamic decision making.
Non technical system Involves Linguistic or Sensory data	Conversational system, Classifier	Regression models, Time series analysis, Decision trees	Root cause analysis systems, Clustering system	Language-controlled manipulation, Object searching	Collaborative agent, Football robot, Self-driving autopilot. Drones
		AI SYSTEMS			AUTONOMOUS SYSTEMS

- AI enables the construction of technical and non-technical systems that would be impossible to achieve with ICT alone!

Validation of Technical 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 formal models for autonomous driving systems, limits their 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.

See: https://ieeexplore.ieee.org/document/11703751
Rigorous Simulation-based Testing for Autonomous Driving Systems -- Targeting the Achilles' Heel of Four Open Autopilots

- We sorely lack testing methods for AI systems similar to those applied to software and hardware systems.
  - **Sampling theory:** methods for constructing samples that adequately cover real-world situations.
  - **Repeatability:** for two samples with the same degree of coverage, the estimated confidence levels are approximately the same.

Even in this case, it is impossible to obtain reliability guarantees of the order of 10^{-8} failures per hour.

Validation of Non-technical AI Systems – Alignment with Human Values



WILL KNIGHT BUSINESS JUN 30, 2025 9:00 AM

Microsoft Says Its New AI System Diagnosed Patients 4 Times More Accurately Than Human Doctors

- Does that mean the LLM should be allowed to practice as a medical doctor?

Certainly NOT! We trust humans because we know they are aware of their obligations and the risks if they fail to comply.

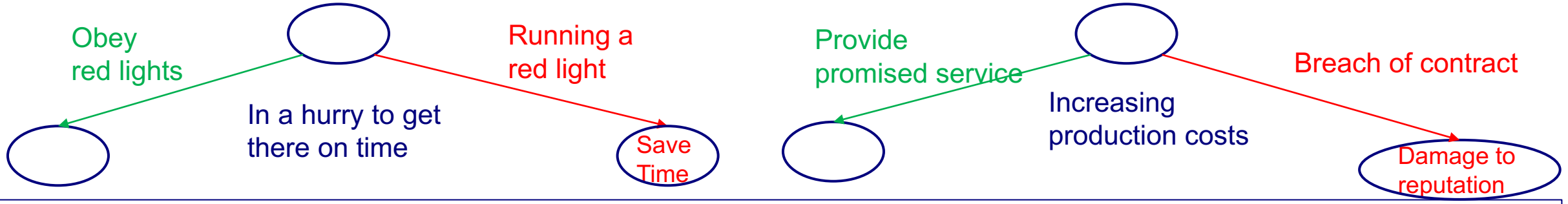
System operation aspect / Requirements	Behavioral Properties		Cognitive Properties
Risk-related properties	Safety	Security	Normative (ethical/law-enforced)
Usefulness properties	Functionality, Performance, Efficiency, User-friendliness		Intent–directed (rational)

- ❑ The comparison between human agents and AI agents must consider
  - **behavioral properties** characterizing the interaction patterns observed between the agent and the world;
  - **cognitive properties** depend on the agent's knowledge, particularly its value system.
    - **Normative properties** characterize the degree of satisfaction of normative rules restricting agent choices;
    - **Intent-directed properties** characterize the way the agent chooses its goals and acts for their satisfaction.

Validation of Non-technical AI Systems – Normative Rational Behavior

- An agent has a set of norms (value system and scales of values) that enable him to estimate the costs and benefits resulting from the execution of actions for himself and his environment. **Normative behavior** implies that:
  - **The agent is aware of conflicting actions** and can assess the impact of its actions;
  - **The agent makes the choice** the most in line with the norms and scales of the value system.

- Unlike behavioral properties, **normative properties** cannot be decided without having access to the agent's world knowledge:
  - saying that "the earth is flat" can be a lie or ignorance;
  - non awareness that I am doing something wrong does not imply my responsibility.



- **Rationality** is a distinctive feature of human thinking that covers a variety of intent-goal-directed properties, such as
  - **Optimal decision-making** and choice (quantitative reasoning);
  - **Coherence**, i.e. problems posed in logically equivalent situations admit similar solutions. (similarity of situations);
  - **Competence levels:** passing a test at a certain level implies passing a test at a lower level! (analogical thinking).

- Rationality simplifies the understanding, analysis and validation of an agent's properties.

  Experimental results show that AI agents are not rational, which makes their testing problem unsolvable!

- Building Multi-Agent Systems
- Validation of AI Systems
- The Way Forward

OVERVIEW

The Way Ahead – Rigorous Intelligent Systems Engineering

- The development of autonomous systems requires **a marriage between AI and traditional systems**, which poses non-trivial technical problems, as new trends are disrupting systems engineering.
  - We need to develop **hybrid reference architectures** that incorporate both machine learning-based components and software components, and that guarantee certain essential properties by construction.
    - How compensate for the lack of predictability through knowledge-based adaptation?
    - How to move from correctness at design time to correctness at runtime to achieve adaptation?
    - How to link symbolic and data-based knowledge e.g. sensory information and models used for decision-making.

- **AI system validation** is marked by the **shift from rationalism to empiricism** due AI's lack of explainability.
  - When a self-driving car is safe enough? We need **rigorous testing techniques** for technical AI systems.
  - In-depth study of cognitive properties and methods for validating them.
  - Develop ontology-based semi-formal methods for semantic text analysis bridging the gap unstructured/structured data.
  - Ensure the determinacy of LLMs through prompt engineering.

- We need to elaborate **a broad technical vision for multi-agent systems** covering a wide range of agent types and their coordination, operating in physical and human environments. This means moving:
  - **from** narrow superintelligence that outperforms humans in performing a task in digital environments—for example, a chess-playing robot that cannot drive a car;
  - **to** broad human-level intelligence, which can only be achieved by combining different types of AI and ICT technologies, including symbolic AI, traditional ML, and LLMs.

(c) J. Sifakis, 2026

Thank you