Shedding Light on the Foggy AI Landscapes

BAAI Conference
Beijing, June 6, 2025

Joseph Sifakis

Verimag Laboratory
and SUSTech/RITAS

Where We Are? – The AI Vision and Its Goals

At present, there's a great deal of confusion about the ultimate goal of AI, fuelled by the media and Big tech companies, who, through grandiose large-scale projects, spread opinions suggesting that human-level AI is only a matter of years away.

Opinions are divided between two very different positions.

- Some AI research and companies such as OpenAI and DeepMind see AGI, as the ultimate goal ***" AGI is a system that can perform any intellectual task a human can, at human-level or beyond*** " S. Altman Feb. 2025
  - AGI definitions are misleading because they implicitly assume that human intelligence can be defined as the ability to perform an undefined set of tasks separately. Which tasks, exactly how many tasks?
  - AGI definitions suggest that it can be achieved through ML and its further developments i.e ML is the "end of the story" Scaling up model size (parameters, data, compute) inevitably would lead to higher intelligence.

It is surprising that such a poorly defined vision has been blindly adopted by experts who, even take the risk of making predictions about its imminent arrival (it's already here, tomorrow, in five years...).

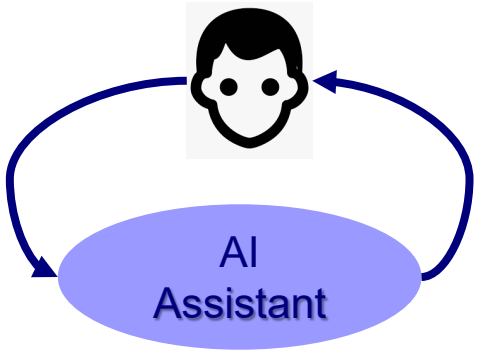
- Others see the goal of AI as building machines with Human-level Intelligence, which requires agreement on what human intelligence is and, more importantly, on methods for comparing human and machine intelligence.
  - According to the Oxford dictionary, intelligence is defined as
    "the ability to learn, understand and think in a logical way about things; the ability to do this well"
  - Machines can perform impressive tasks, outperforming humans in their execution, but there is no evidence that they can surpass humans in terms of 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!

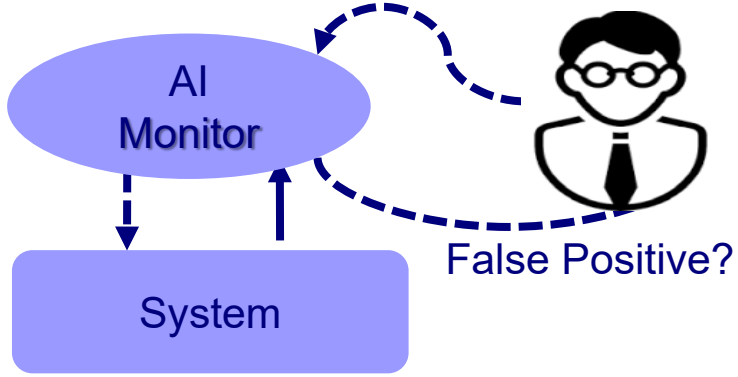
Where We Are? – From Conversational to Autonomous AI

- 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 mainly focuses on assistants, while future applications require continuous interaction with little or no human intervention.

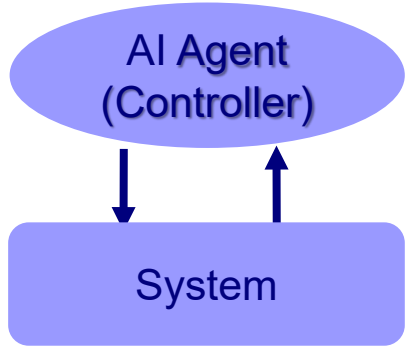
- Three different ways to use AI systems:
  1. Assistants that in interaction with a user, provide a given service;
  2. Monitors of a system behavior synthesizing knowledge to detect or predict critical situations;
  3. Agents (controllers) of a system so that its behavior meets a given set of requirements,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



Conversational AI



Predictive/Analytical AI



Autonomous AI

- The impact of AI on the real economy remains limited compared to the enormous potential of AI applications
  - The AI industry revolution has only just begun!
  - Its realisation depends largely on our ability to develop AI-enabled agents to build autonomous systems.

Where We Are? – Technical vs. Non-technical Systems

- Technical systems: their I/O relation can be unambiguously characterized as a relationship between mathematical domains.

  Traditional ICT systems are technical systems
  - Their trustworthiness can be formulated as predicates P(x,y) that can be validated by testing.
- Non-technical systems: their input or output domains are defined in terms of sensory or linguistic data.
  - These are systems that mimic human functions not amenable to formalization, e.g. ChatGPT, image analysis system.
  - Their behavioral properties cannot be validated as rigorously as those of technical systems., e.g, safe GPT.
  - Their trustworthiness encompasses human-centric properties determining their degree of alignment 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.!

Where We Are? – The Vision for Autonomous and Trustworthy AI

- Autonomous systems stem from the need to replace human operators in complex organizations as envisioned by the IoT e.g, self-driving systems, smart grids, smart factories, autonomous telecommunication systems
  - are composed of agents each pursuing their own goals (individual intelligence) while coordinating to meet global system goals (collective intelligence);
  - are able to perceive and predict changes of their environment and adapt to the constantly changing environments and user requirements by managing possible conflicting goals;
  - are highly complex, often critical distributed dynamic reconfigurable systems that never stop and evolve.

- A key issue concerning the extensive use of AI systems, reputed to be black boxes, is to guarantee their trustworthiness.
  - Safe AI has been the subject of international summits and UN deliberations, marked by naive optimism and ignorance of the fact that LLM safety cannot be guaranteed in the same way as that of an elevator, an airplane, or a car.
  - Human-centric cognitive properties, in addition to behavioral properties, are the subject of numerous studies.
    - "Responsible AI" implies that AI systems meet properties such as fairness, reliability, safety, privacy and security, inclusiveness, transparency, and accountability, difficult, if not impossible, to assess.
    - "AI alignment" meaning alignment with human ethical values while we do not even understand how human will emerges and the associated value-based decision-making system works.

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.*

Overview

- Agent Reference Architecture
- Validation of AI Systems
- Where Are We Going?

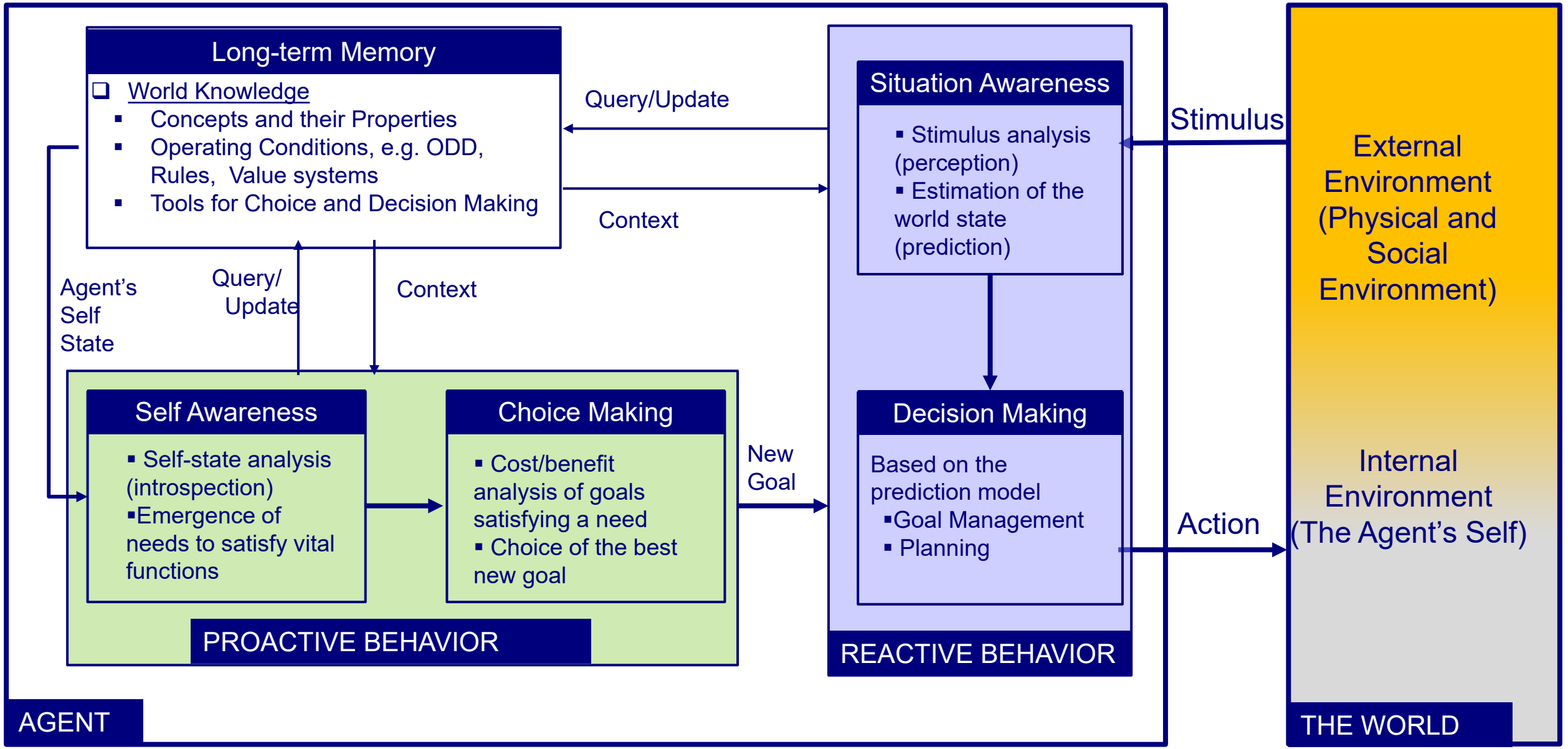
Agent Reference Architecture – Principles

- With the advent of LLMs, AI agents have attracted growing interest,
  - used to provide services mainly by interacting with static digital worlds, e.g. business processes, problem solving, which is insufficient to cover the needs for autonomous agents
  - built based on architectures integrating LLMs augmented with knowledge generated by engines or stored in a Memory
    - LLMs grounded to symbolic engines e.g. AlphaGeometry, WolframAlpha, simulators, probabilistic programming tools…
    - LLMs use World Knowledge stored in long-term memory, e.g. Retrieval-Augmented Generation (RAG).

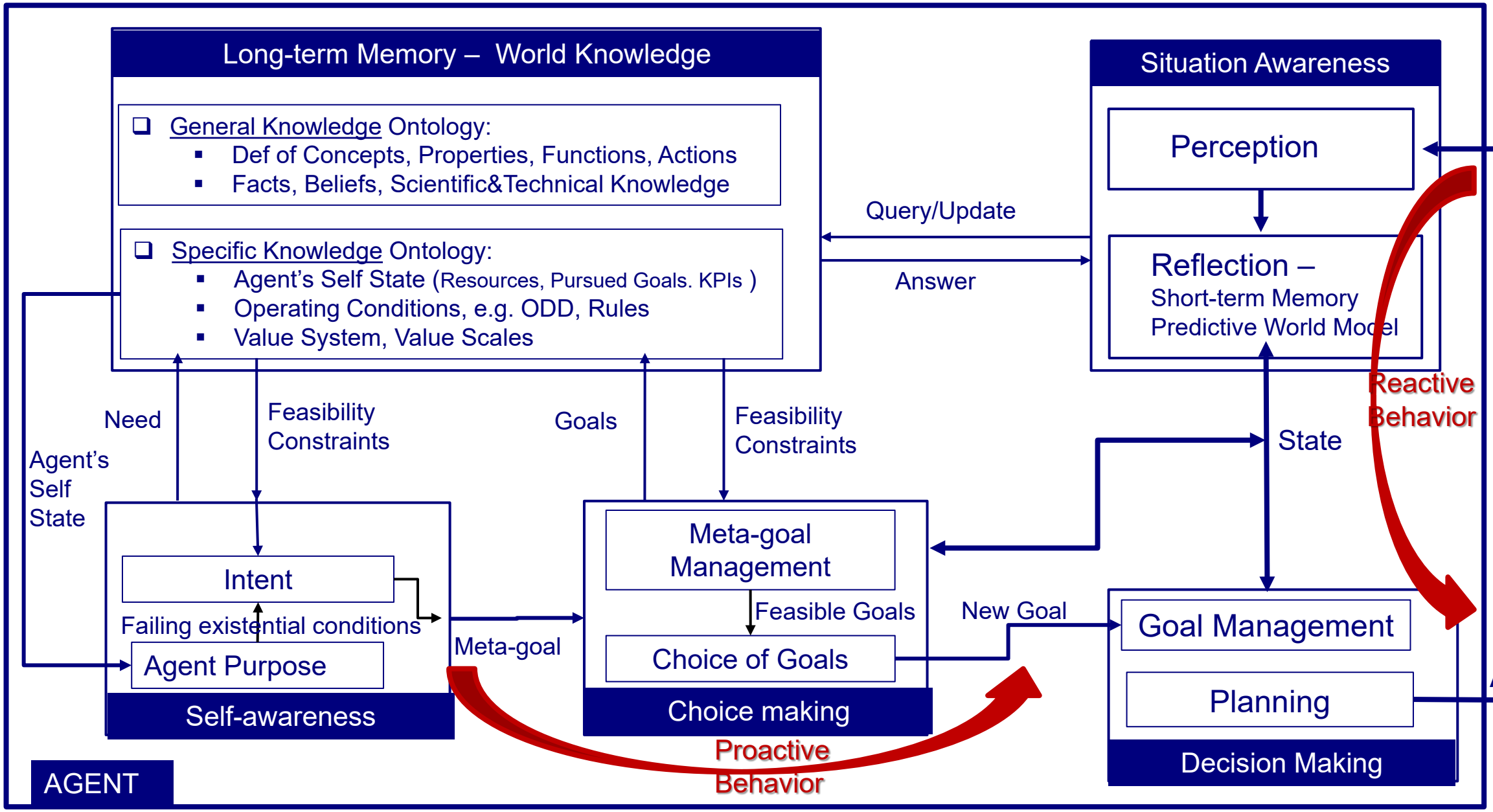
- We need an Agent Reference Architecture generalizing existing solutions, designed to capture cognitive human behavior, and which could serve as a benchmark for evaluating agent implementations.
  - Integrates a set of basic functions, mathematically definable and independent, assuming that intelligence is an emerging property of computation.
  - Shows sufficiently general and complete behavior, covering as far as possible various aspects of human intelligence, in particular Kahneman's two systems of thinking.

- An agent Reference Architecture could serve as a basis for
  - characterizing autonomy as the composition of functions from which can emerge an intelligent behavior;
  - the comparison of various agent solutions and their degree of coverage of model features
    - Ranging from end-to-end machine learning to hybrid architectures.
    - Targeting specific application domains such as ADS, Robotics, business decision support, gaming, etc.
  - addressing agent correctness through compositional reasoning and separate validation of the agent's constituent elements.

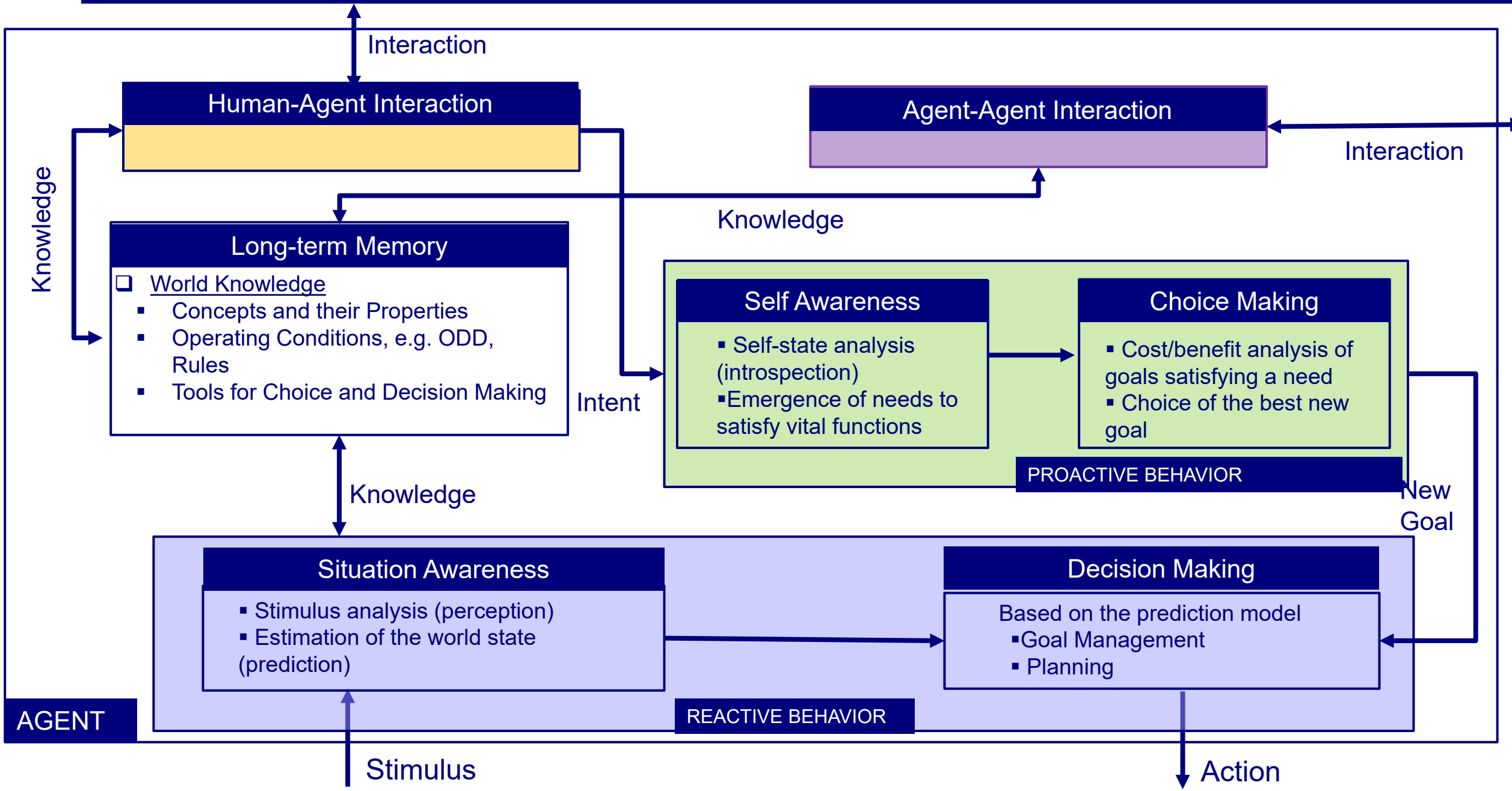
Agent Reference Architecture – General View



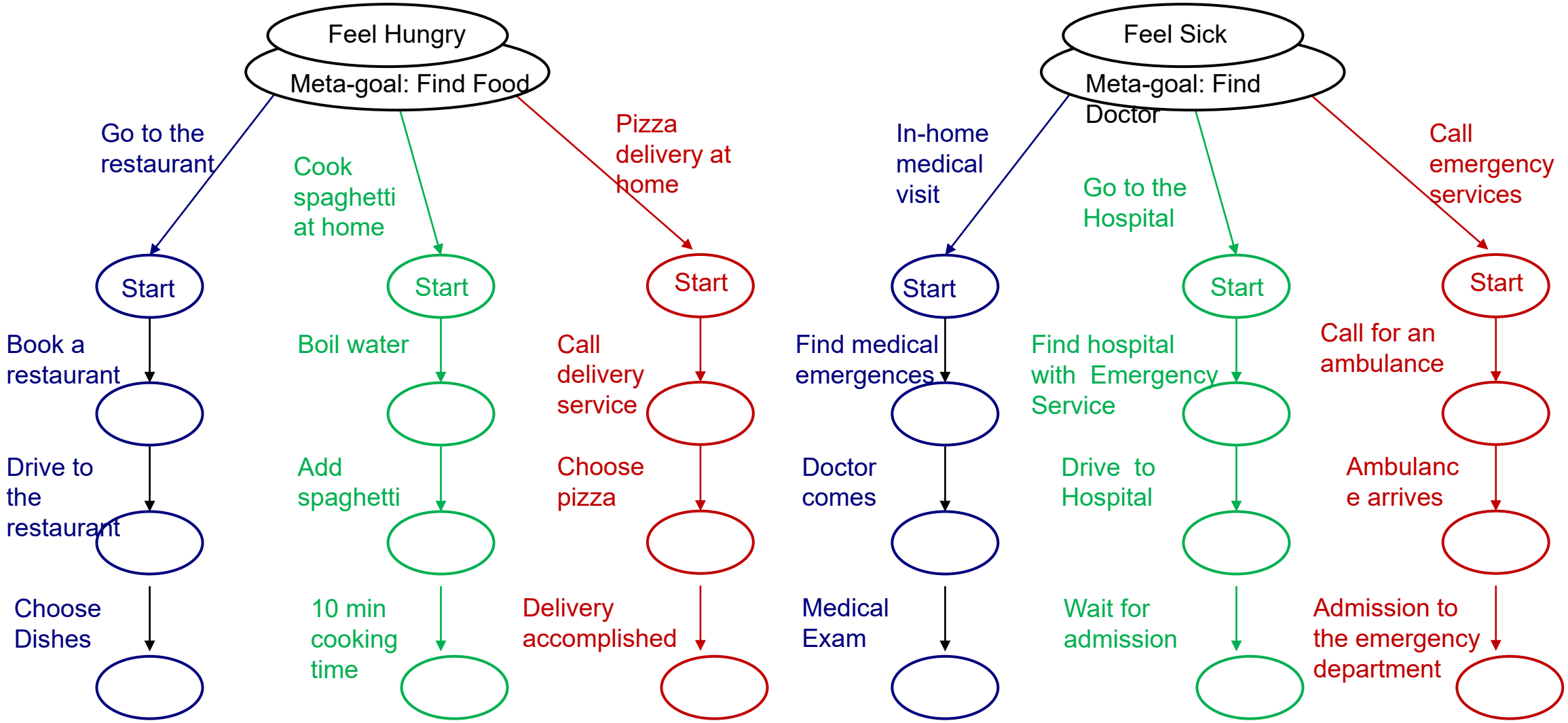
Agent Reference Architecture – Detailed View (1)



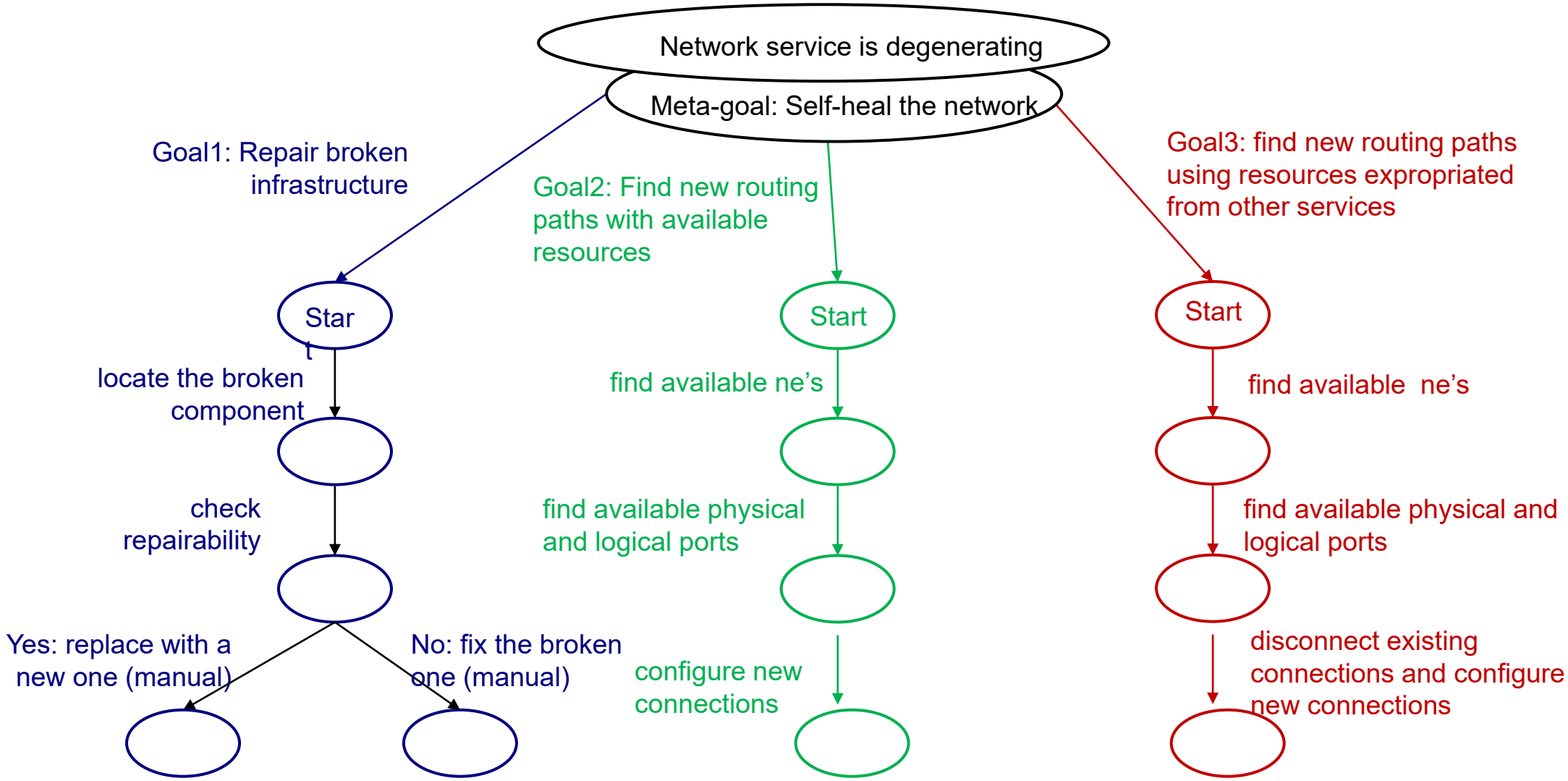
Agent Reference Architecture – Communicating Agent



Agent Reference Architecture – Proactive Behavior: Meta-goals and Goals



Agent Reference Architecture – Proactive Beahvior: Meta-goals and Goals



Source: Dongming Li et al. Huawei

Agent Reference Architecture – Two Key Challenges

- 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.
  - Planning for the achievement of goals is a two-player game between the agent and its environment played on the state space of a predictive model describing their interactions from their initial states, a lookahead tree.
    - Explicitly building the lookahead tree involves exponential complexity – only when the rules of the game are well-defined, we can use Monte Carlo Tree Search (MCTS) algorithms as in AlphGo.
    - Control policies are driven by goals involving 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.

- Ontology-based knowledge management in architectures organized around a long-term memory containing both symbolic and non-symbolic knowledge to bridge the gap between the world of knowledge and the world of language.
  - Embedding techniques prove to be largely non sufficient to account for semantic subtleties.
  - We should develop knowledge graph technology based on hypergraphs that represent n-ary relations between entities.
  - We should develop domain-specific parameterized ontologies that can be linked to higher-order logic languages.
  - We should develop retrieval mechanisms based on specific similarity relations including perceptual, semantic, thematic, and functional.,
  - We should develop consistency checkers for knowledge management, e.g. after updating knowledge of the memory.

Agent Reference Architecture – World Knowledge

Knowledge

Agent-dependent (subjective) Knowledge: Cognitive and Deontic

Agent-independent (intersubjective) Knowledge: Scientific and Technical

Declarative
Implicit empirical
Mathematical and Scientific Theories
Properties, Declarative rules

States – State Predicates

Procedural
Empirical Methods
Design and System development flows
Coordination principles, Imperative rules

Actions – Action Predicates

Implicit human knowledge
Machine learning knowledge

Data-based

Model-based

Explainable by models e.g.
formal or simply conceptual

Objects, Physical or Artefacts
Agents, natural or artificial

Concrete

Concepts

Abstract

A posteriori e.g. Scientific and Technical
A priori e.g. Mathematics, Logic
Self-concepts e.g. Intention, Belief, Feeling

Types of knowledge, depending on their degree of validity, generality and domain and mode of use

Agent Reference Architecture – Linking LLM to Ontologies

WORLD KNOWLEDGE

- Basic knowledge about the world characterizes the world states that satisfy atomic predicates P(x,y, ….).
- Temporal knowledge about the world characterizes state sequences *seq= state1 state2, .., state_n* using formulas with quantification over sequence and their states
  - *always P(x,y,..) = ∀seq∈SEQ ∀i P(seq(i)(x,y,…))*
  - *inevitable P(x,y,..) = ∀seq∈SEQ ∃i P(seq(i)(x,y,…))*
- Spatial knowledge about the world characterizes relations between positions of the entities in a space: "*a follows b*" means that *distance(b.pos(t)-a.pos(t)) ≤ d for all t*.
- Epistemic knowledge about the world uses the modality k_x to express the fact that agent *x* knows a property *P* e.g. $k_x(P)$
- Deontic knowledge about obligations, permissions to perform actions.

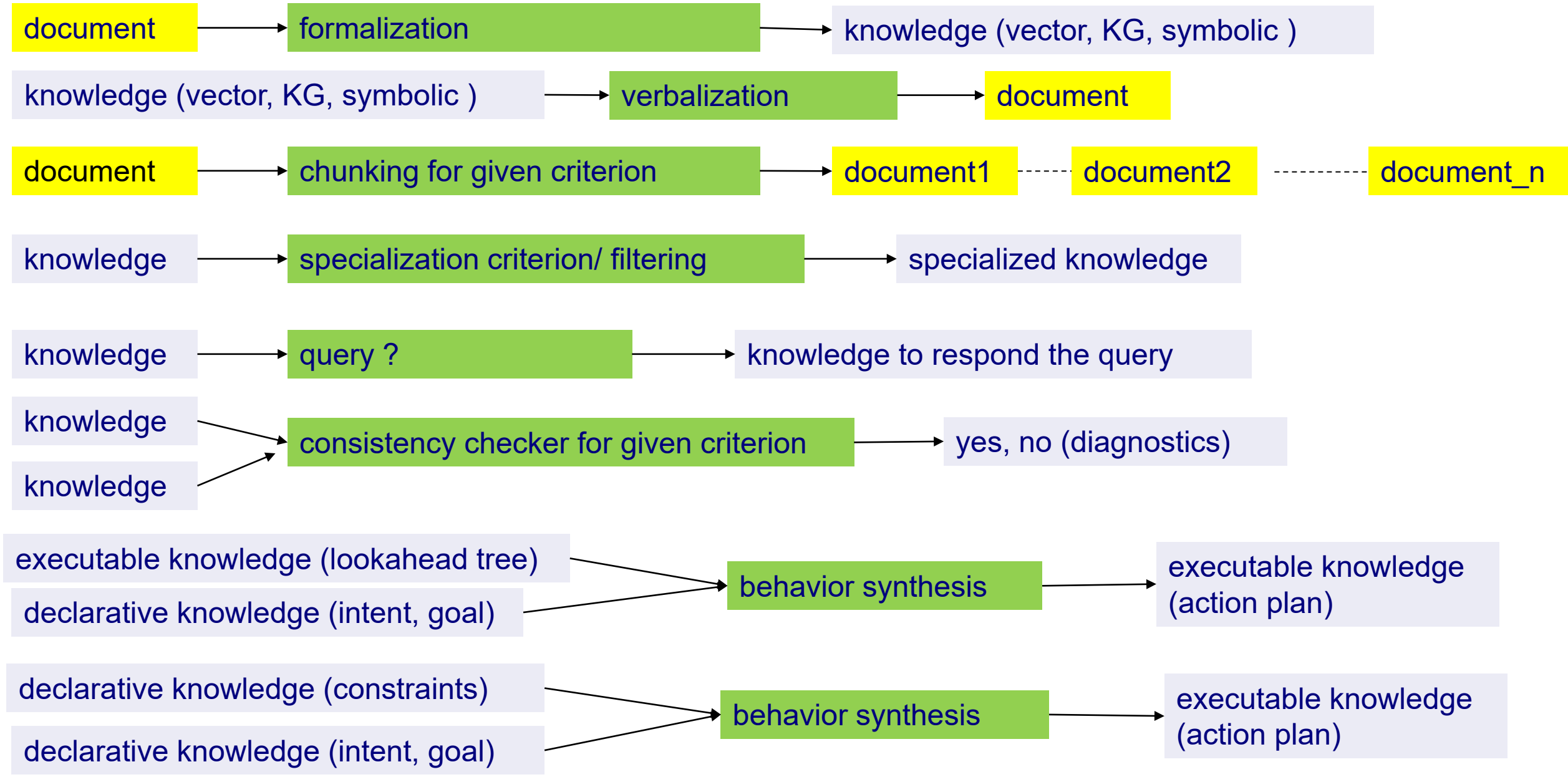
Verbalization

Formalization

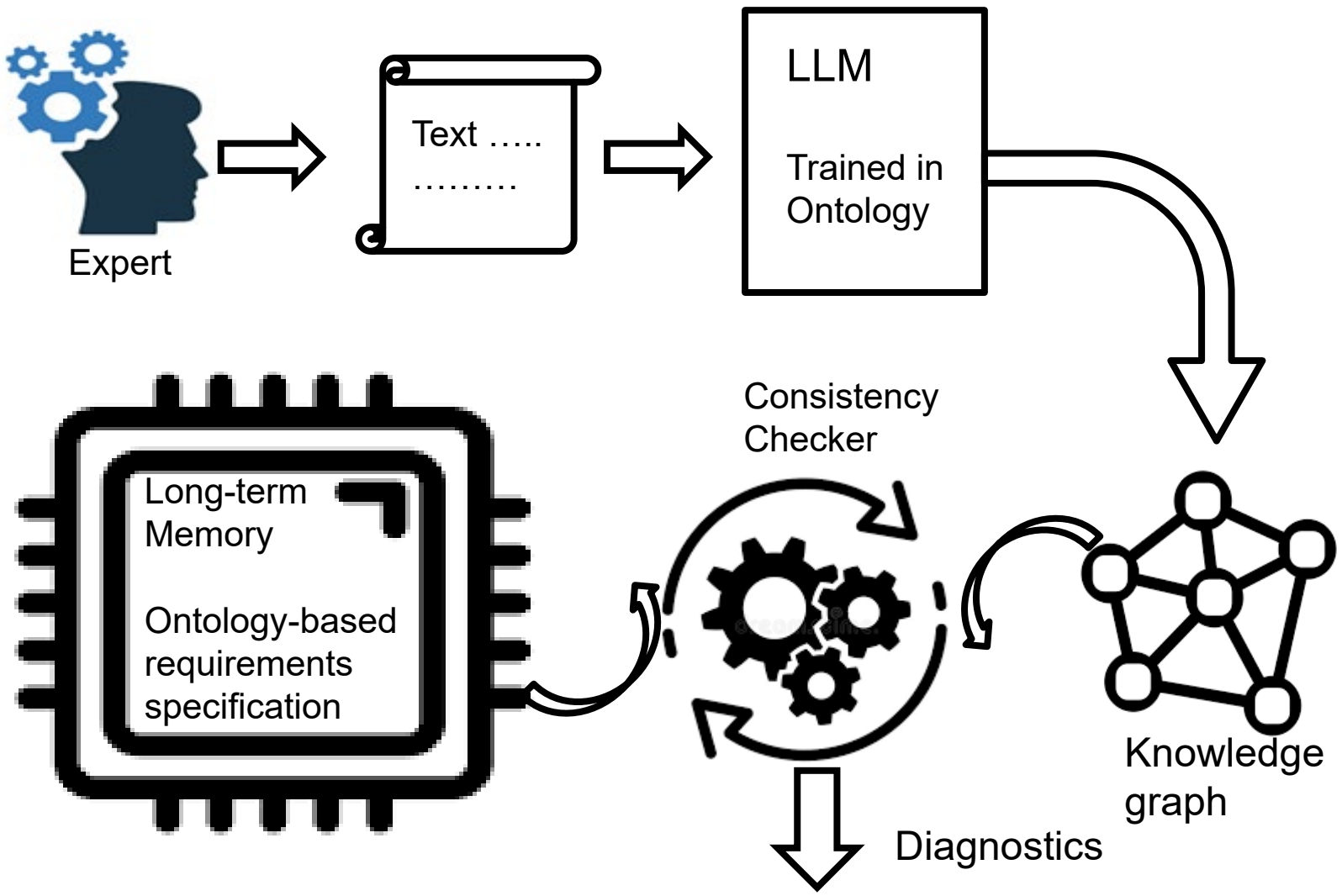
NATURAL LANGUAGE

- Establish correspondence between Ontologies and Natural Language requires specific tokenization techniques :
  - Contexts characterizing states or sets of states of the world as relationships between concepts and their attributes involving "is" and "has".
  - Actions that correspond to change of contexts expressed by verbs denoting change, intention to do with two modalities: do(action) and say(text).
  - Temporal modalities, such as always, eventually, possibly, ever, maybe, may, might, after, before ….
  - Space modalities such as above, below, left, right, follows, precedes, between, containment relation.
  - Epistemic modalities, such as know, believe, think, …
  - Deontic modalities, such as must, have to, obliged to, mandatory that, imposed on,

Agent Reference Architecture – Additional Problems To Be Solved



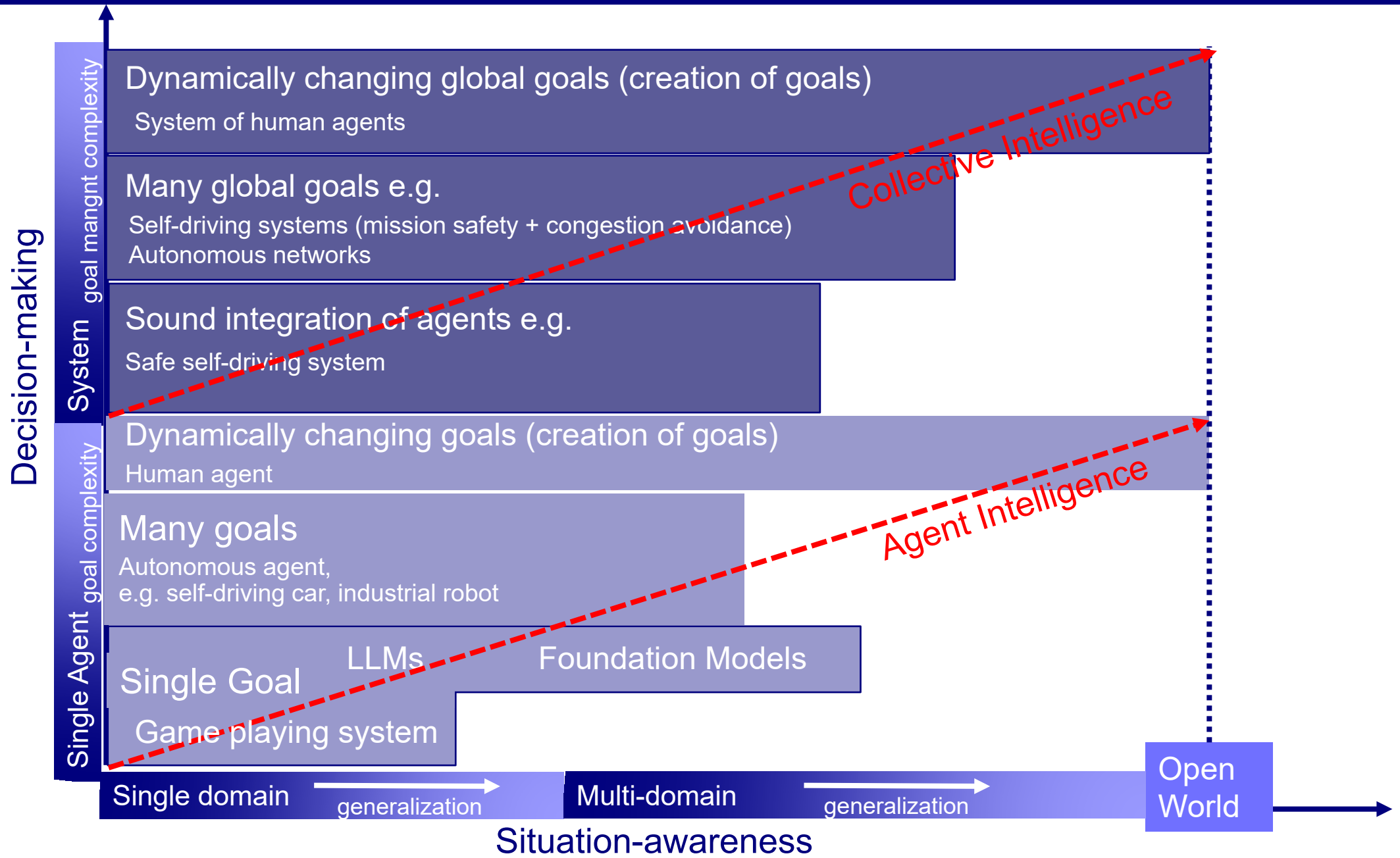
Agent Reference Architecture – Consistency Checking (Safeguarded AI)



Use case :

- An O&M engineer writes a scenario describing the steps for configuring an autonomous network.
  Before applying the scenario, they want to ensure that the generated configuration will not affect the essential requirements of the autonomous network, ranging from connectivity to dynamic load balancing, energy efficiency, and quality of service (QoS) guarantees.
- The long term memory contains an ontology-based requirements specification for ANs
- Use a Consistency Checker to compare the knowledge graphs with the ontology stored in a Memory and generate validation results, possibly diagnostics pointing out inconsistency.

Agent Reference Architecture – From Agent to Collective Intelligence



Overview

- Agent Reference Architecture
- Validation of AI Systems
- Where Are We Going?

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 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.

- 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 of operation required for critical systems.

Validation of AI Systems – Alignment with Human Values

- What do artificial agents lack to approach the characteristics of human behavior?
  - Can an agent be made credible in such a way that it gives the impression of human behavior?
  - What are the distinctive human features that machines have difficulty reproducing?
- Many consider that outperforming humans in behavioral Q&A tests is proof of machine intelligence. But is it enough?
  - There is every reason to believe that an LLM will be able to pass the final medical exams as successfully as the students.
  - Does that mean the LLM should be allowed to practice as a medical doctor?

Certainly NO! We trust humans because we know that they know the rules of a value system and are bound by them.

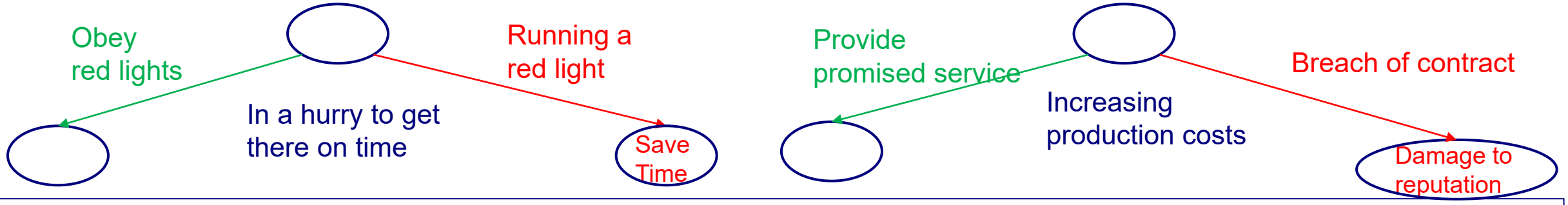
System operation aspect / Requirements	Behavioral Properties		Cognitive Properties
Risk-related properties	Safety	Security	Normative (ethical/law-enforced)
Usefulness properties	Functionality, Performance, Efficiency, User-friendliness		Intent and Goal-directed, Rationality properties

- 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 and Goal-directed properties characterize the way the agent chooses its goals and acts for their satisfaction.

Validation of AI Systems – Acting Ethically and Rationally

- An agent is equipped with a value system, a set or rules and value scales for estimating costs and benefits of its actions.
  It acts ethically if
  - It is aware of conflicting actions and can assess costs/benefits of actions for itself and for the agents in its world
  - It makes the choice the most in line with the rules of the value system.

- Unlike behavioral properties, ethical 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 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!

Validation of AI Systems – Formalizing Cognitive Properties

- Consider formulas built from the set of atomic predicates below, where x,y are agents, p is knowledge, and α is an action.
  - *do_x(α)*: agent x executes action α;
  - *say_x(p)*: agent x asserts that p is true;
  - *k_x(p):* agent x knows (believes) that p is true.
  - *vl_y(do_x(α)):* value generated for agent y, according to its value system, by action α executed by x,
  - *wr_y(do_x(α)):* agent y considers it incorrect (contrary to its normative rules) for x to perform action α;

- The following normative properties can be expressed using the above predicates:
  - Dishonest: *say_x(p) and k_x(not p)*
  - Irresponsible*: do_x(α) and k_x(wr_x(α))* *// x performs action α knowing that it is forbidden;*
    Not responsible: do_x(α) and not(k_x(wr_x(α))) // x performs *α* without knowing that it is incorrect.
  - Selfish: *do_x(α) and k_x(wr_x(α)) and vl_x(do_x(α))>>0) and vl_y(do_x(α))<<0)*
    //x knowingly performs wrong action α, beneficial to him and detrimental to y.
  - Generous: *do_x(α) and vl_x(do_x(α))<0) and vl_y(do_x(α))>>0)* *// x performs α, detrimental to him, but beneficial to y.*
  - Stupid: *do_x(α) and vl_x(do_x(α))<0) and v_y(do_x(α))<0)* *// x performs action α detrimental to him and to y*
  - *Trust: k_x(vl_y(do_y(α)) > vl_y((vl_y(not do_y(α))) // x trusts y to do action α because x believes it is beneficial to y.*

- The validation of cognitive properties presupposes the evaluation of atomic predicates on the agent's knowledge,

Overview

- Agent Reference Architecture
- Validation of AI Systems
- Where Are We Going?

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, as new trends are disrupting traditional systems engineering.
  - How can reliable systems be built from unreliable components using hybrid architectures that integrate ICT components and unexplainable AI components, while getting the best out of each?
  - How to link symbolic and non-symbolic knowledge e.g. sensory information and models used for decision-making.
  - How to move from correctness at design time to correctness at runtime to achieve adaptation?

- System validation is marked by an irreversible shift from rationalism to empiricism due AI's lack of explainability.
  - We must strive to compensate for the lack of solid guarantees of trustworthiness by using explicit knowledge about the world stored in long-term memory.
  - We need technical standards that provide methods for risk assessment and reliability certification.
    - Standards do not hinder innovation; on the contrary, they challenge us to find innovative solutions.
    - Absence of regulation leads to poorly engineered systems, increase technical debt that compromises the future.

- We need to elaborate a broad technical vision covering a wide range of system types and domain-specific technologies.
  - Human intelligence has many facets and can only be achieved by combining different types of AI and ICT technologies, including symbolic, traditional ML and LLM, e.g. a chess playing robot, cannot drive a car.
  - The setbacks experienced by the autonomous car industry show that there is still a long way to go to bridge the gap between automation and autonomy.

Where Are We Going? – Multi-agent Systems

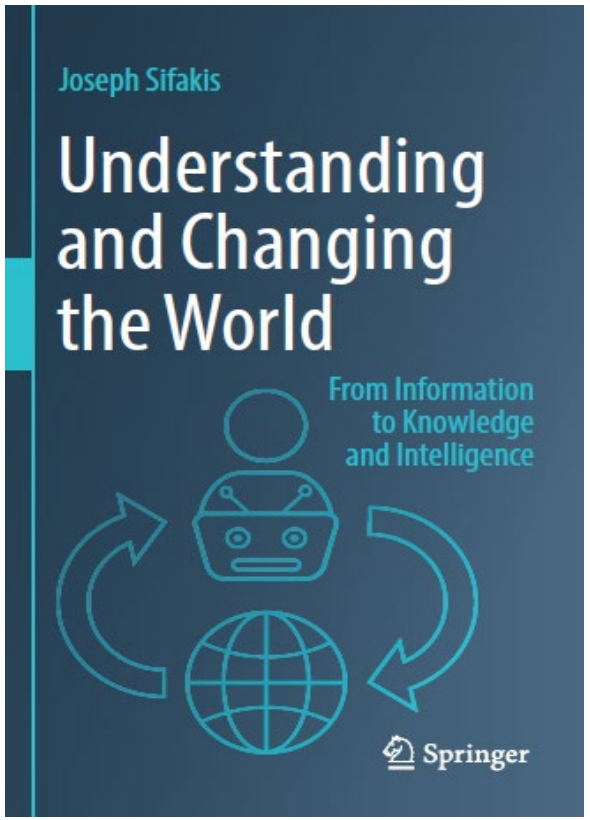
Multi-agent systems require convergence between three distinct areas of knowledge

- ❑ Traditional systems and software engineering technology
  - Existing results from distributed systems theory, in particular algorithms for achiveng resilience, consensus
  - Communication protocols that support various modes of coordination and dynamic configuration in order to adapt to failures and a dynamically changing environment
- ❑ Results from traditional multi-agent systems consisting of SW agents which
  - rely on symbolic local knowledge and the use of rule based systems to make decisions
  - can be studied using BDI logic frameworks including epistemic, deontic and dynamic logics
- ❑ AI to deal with data-based knowledge and in particular with linguistic and sensory data.

Is such a convergence possible?

- ❑ A look at the A2A protocol, a framework for creating MAS, says a lot about the challenges involved in meeting requirements such as, 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 aimed at ensuring essential cognitive properties such as trust, rationality, and accountability.
  - These are known problems whose difficulty is already recognized in symbolic/model-based frameworks.
  - AI is a game changer, offering greater flexibility in specifications and greater efficiency in problem solving, but at the cost of a serious lack of rigor and semantic control.

The solutions, if feasible, will require a combination of symbolic and data-driven techniques—and that will take hard work.


Joseph Sifakis
Understanding and Changing the World
From Information to Knowledge and Intelligence
Springer

Thank you


图灵奖得主、中美法三国院士重磅著作
理解和改变世界
[法] 约瑟夫·希发基思（Joseph Sifakis）/ 著
唐杰 阮南捷 / 译
从信息到知识与智能
UNDERSTANDING AND CHANGING THE WORLD
看清世界的底层逻辑，以更高维的方式思考和决策
Springer
中信出版集团