Testing System Intelligence – A Systems Engineering Perspective

Joseph Sifakis

Verimag Laboratory

Data Intelligence Institute
Paris
December 6, 2023

Overview

- ❑ Where We Are?
- ❑ Testing System Intelligence
- ❑ Where Are We Going?

Where We Are?

At present, there is a great deal of confusion as to the final objective, with opinions divided between two very different positions reflecting the lack of agreement on what intelligence is:

❑ 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

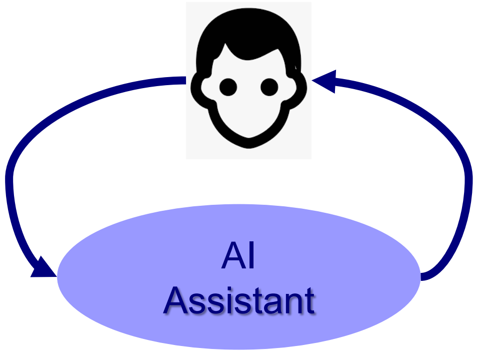
□ Despite the impressive growth in AI we have seen in recent years, culminating in the arrival of generative AI and its application to solving NLP problems that have always remained open, we only have weak AI that

- 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.

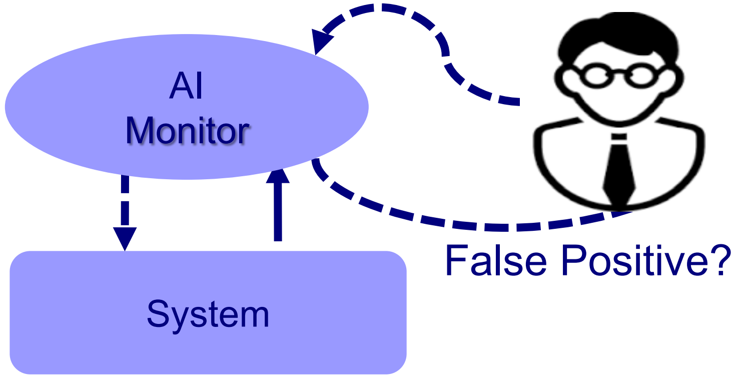
□ There are three different ways to use AI systems :

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.

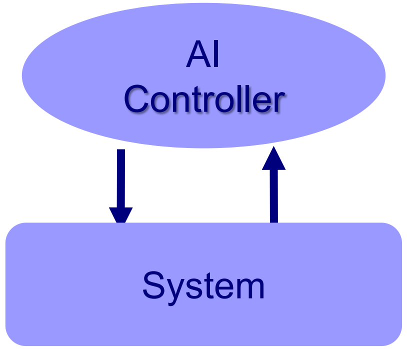
Monitors and Controllers will be by far the most important in the future for building intelligent products and services.



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? – AI Risk Management

- ❑ Managing the risks associated with AI has long been a concern for the authorities and institutions, who, despite numerous consultations and legislative efforts, have so far failed to come up with concrete and realistic regulatory frameworks.

- ❑ Recently, given the stakes involved in adopting common regulations, the United States, the European Union and also China are currently working on regulatory frameworks to manage the risks associated with the use of AI and ensure that society as a whole benefits from its advances. Is world-wide AI governance necessary and possible?

- ❑ The elaborated regulatory frameworks are very different in purpose, scope and approach.
  - ▪ The AI Executive Order (issued on October 30 this year) is not coercive; it includes recommendations and guidelines seeking consensus and collaboration from key US players.
  - ▪ Chinese texts follow a vertical approach regulating three types of systems (recommendation, deep synthesis and generative) impose stringent content constraints that require careful design from the outset.
  - ▪ European regulations define both a rigorous framework for risk management and a set of rules to be respected by the very large online platforms and search engines.
    - ○ The AI Act adopts a risk-based approach distinguishing four levels of risk (unacceptable risk, high risk, limited risk, minimal risk or zero risk), which require corresponding guarantees of reliability.
    - ○ The Digital Services Act regulates online intermediaries and platforms, and defines the obligations with which 19 very large online platforms (VLOPs) or very large online search engines (VLOSEs) must comply by February 2024.

Where We Are? – Requirements for AI Systems

The application of AI regulations implies requirements for the systems developed and deployed.

❑ These include trustworthiness requirements that traditional digital systems must satisfy, e.g.

- safety: the system will not enter critical states that could harm humans or the physical environment;
- security: resilience to attacks by unauthorized users that could compromise data confidentiality, integrity and availability.

❑ In addition, a great deal of work is aimed at building AI systems meeting a set of human-centric ethical principles, such as

- Empathy: understanding the social implications of responses to humans and respecting human emotions and feelings.
- Transparency: The decision-making mechanisms must be clear to promote accountability and scrutiny.
- Fairness: The systems must not violate human rights regarding sex, race, religion, gender, and disability.
- No bias: Training data must be regulated and evaluated to detect and eliminate biases that the data may perpetuate.
- Accountability: Users can determine who is responsible for protecting them against any adverse outcomes from AI

However, all these works lack foundation, because they ignore 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?

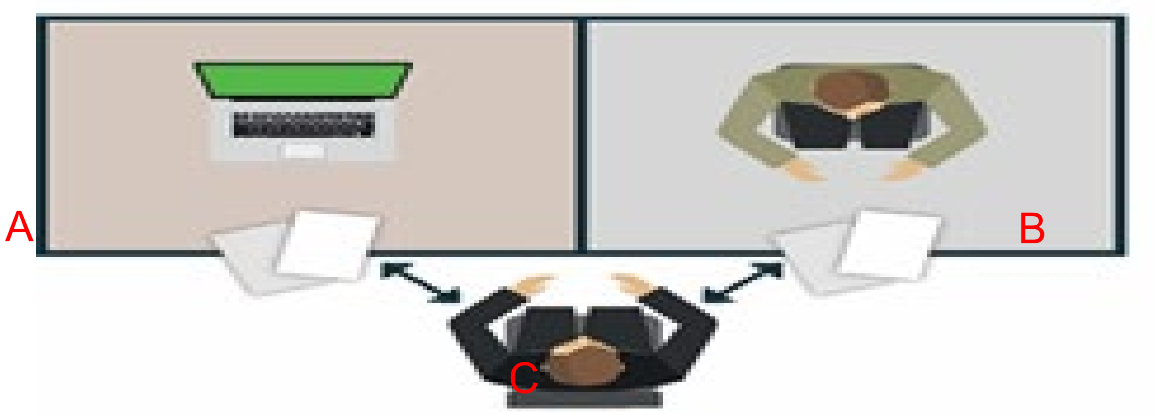
Overview

- ❑ Where We Are?
- ❑ Testing System Intelligence
- ❑ Where Are We Going?

Testing System Intelligence – Behavioral Tests

- ❑ Turing Test (Imitation Game):

1. C sends questions to A and B who, in turn, provide a corresponding answer to each question.
2. If C cannot tell which is the computer and which the person, then A and B are equally intelligent.



- ❑ Criticism:
  - Success depends on human judgement (subjective) and the choice of the test cases (questions).
  - The test cannot be a question/answer game - much of human intelligence is expressed by interaction with the environment (speech, movement, social behavior, etc.)

- ❑ Replacement test: *An agent A (indifferently machine or human) is as intelligent as an agent B performing a given task characterized by given well-founded success criteria, if A can successfully replace B. e.g.*
  - a machine is as intelligent as a human driver is if it can successfully replace the driver.
  - a human is as intelligent as a janitor robot if it can successfully replace the robot according to given cleaning criteria.

Note: The replacement test applies to autonomous systems, while the Turing test is a special case for conversational tasks.

Testing System Intelligence – About Testing

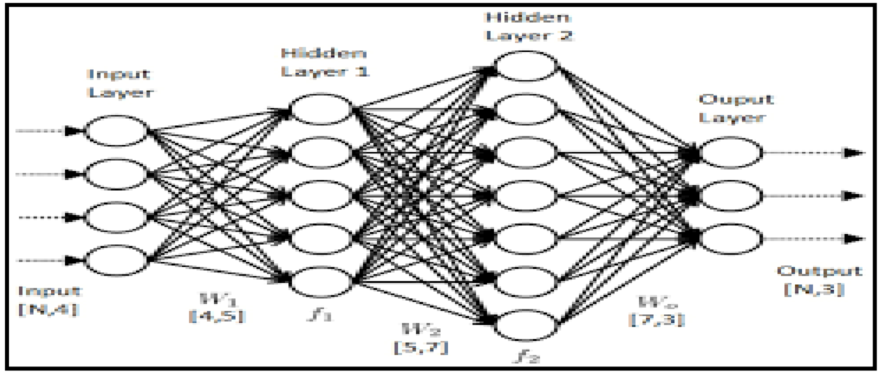
- ❑ Testing behavioral equivalence boils down to testing that a system y= S(x) satisfies a predicate P(x,y) relating stimuli to responses. For example,
  - Turing test: P(x,y) =true if y is a correct answer to question x.
  - Replacement test for self-driving cars: P(x,y)=true if y is a safe trajectory for the driving scenario x.

- ❑ Testing is an essential part of the scientific method applied to the production of empirical knowledge in all disciplines.
  - It differs from verification, which is validation by reasoning on a system model that can assert the validity of universally quantified properties such as safety.
  - 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 confirm property validity; property validity means non-falsification for
    - ○ a very large number of tests e.g. random testing
    - ○ test generation guided by a model of the system behavior e.g. white box testing
  - It is a basic validation technique for traditional digital systems, for which there are a variety of white box test methods, e.g. structural test, functional test, metamorphic test.

To what extent can test methods be applied to intelligent systems?

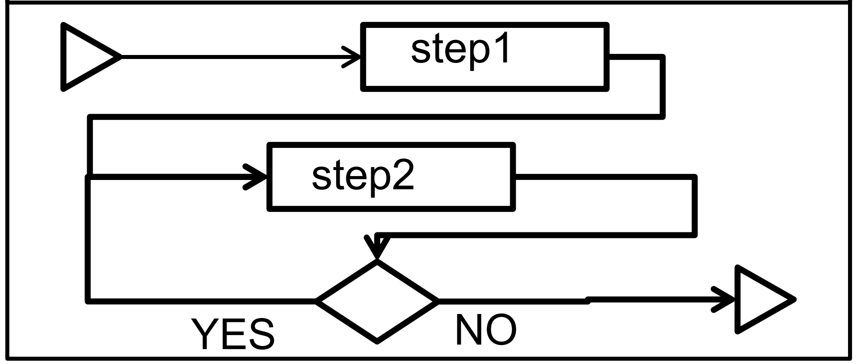
Testing System Intelligence – Significant Differences

NEURAL NETWORK



- Generate empirical knowledge after training (Data-based knowledge)
- Are “black-box” not explainable.

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.

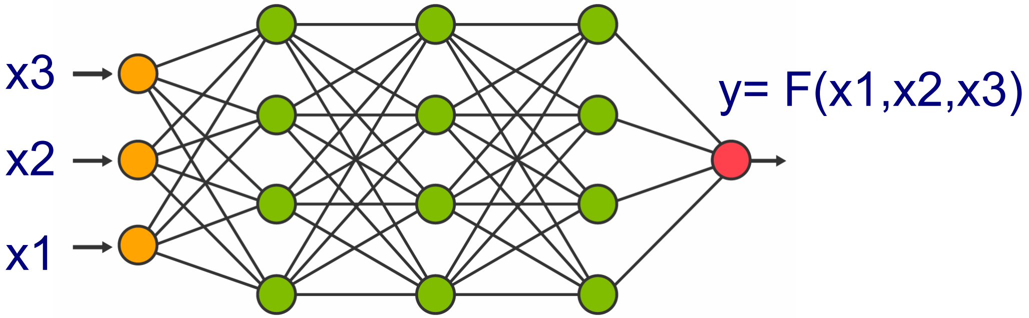
- Neural Networks do not execute algorithms, we use algorithms to train them!

- 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.

Testing System Intelligence – AI Explainability

- ❑ A system is explainable if its behavior can be described by a model that lends itself to reasoning and analysis. 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.



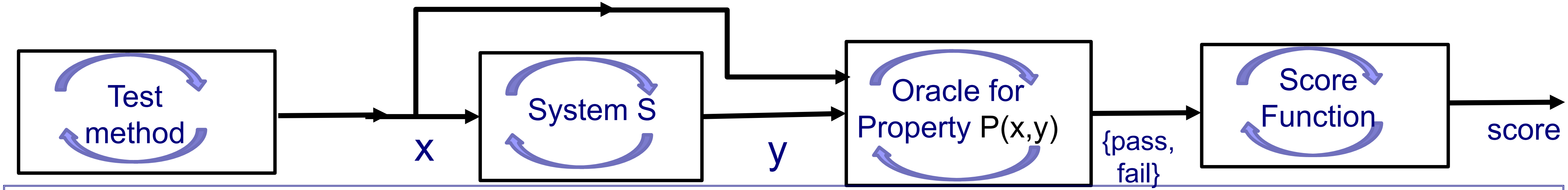
- For feed-forward networks, it is theoretically possible to calculate the output as a function F of the inputs, given the functions calculated by each node: *φ(weighted_sum_of_inputs),* where φ is an activation function.
- However, the approach does not scale up for NN's in real-life applications. Only 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 properties, e.g., extracting a textual description of behavior or decomposing into informally specified elements.

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. a physical system, artifacts like autopilots and AI components;
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 to choose among the 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:

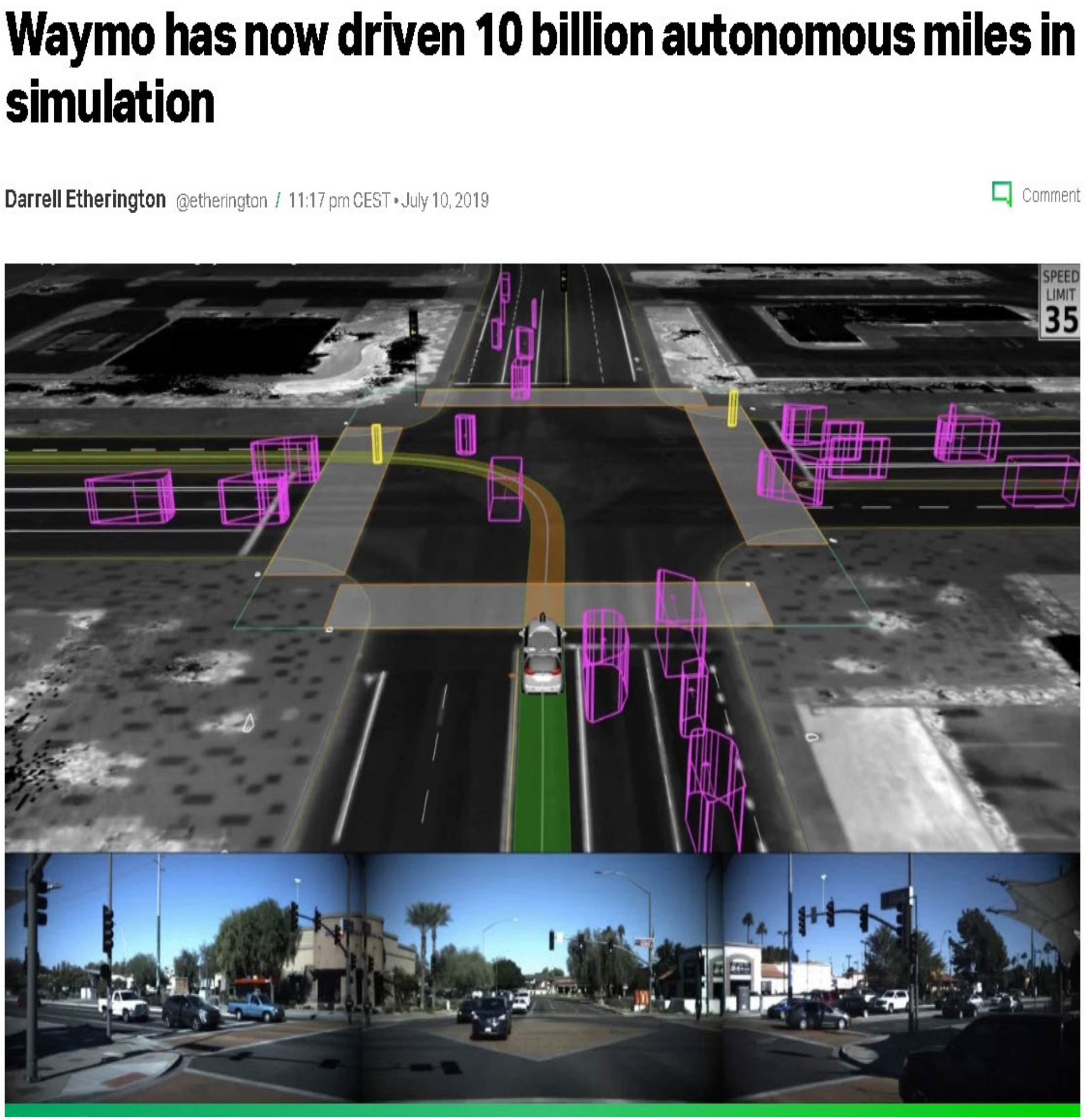
coverage(X1)=*coverage*(X2) implies *score*(X1,Y1) ~ *score*(X2,Y2*)*

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 properties of technical systems 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 Self-driving Cars are Safe Enough?



Waymo has now driven 10 billion autonomous miles in simulation

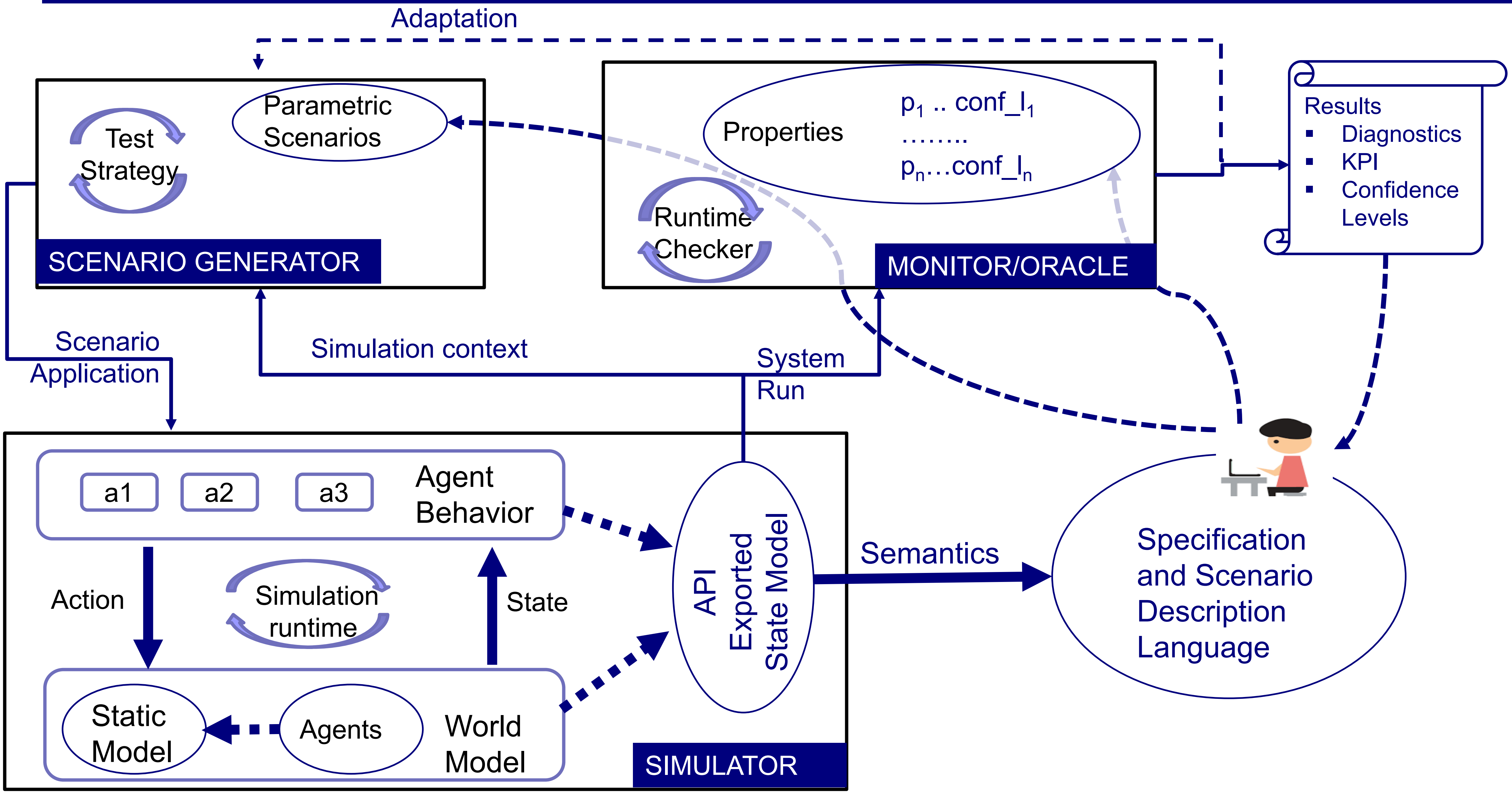
Darrell Etherington @etherington / 11:17 pm CEST • July 10, 2019

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.

Validation of Intelligent Systems – Simulation and Testing Environment



Validation of Intelligent Systems – In Quest of Foundation

❑ We need validation theory inspired from existing white box testing methods where scenarios play the role of test cases allowing the controlled execution of agents so as to explore critical situations based on

- coverage criteria measuring the degree to which relevant system configurations have been explored, as for structural testing of software systems;
- functional criteria to explore/detect corner cases and high risk situations, exactly as for functional testing software systems;
- verdicts and diagnostics about the relationship between failures and various risk factors.

❑ We are currently developing a test method for autonomous vehicles based on scenario compositionality and classification.

- The test for a given scenario can be broken down into a set of scenario types corresponding to different types of maneuver.
- For each type of scenario, sets of the most dangerous configurations can be calculated.

The experimental results obtained with Carla/Apollo reveal considerable flaws that are highly unlikely to be detected by random testing.

Overview

- ❑ Where We Are?
- ❑ Testing System Intelligence
- ❑ Where Are We Going?

Where Are We Going? – The Space of Possible Intelligences

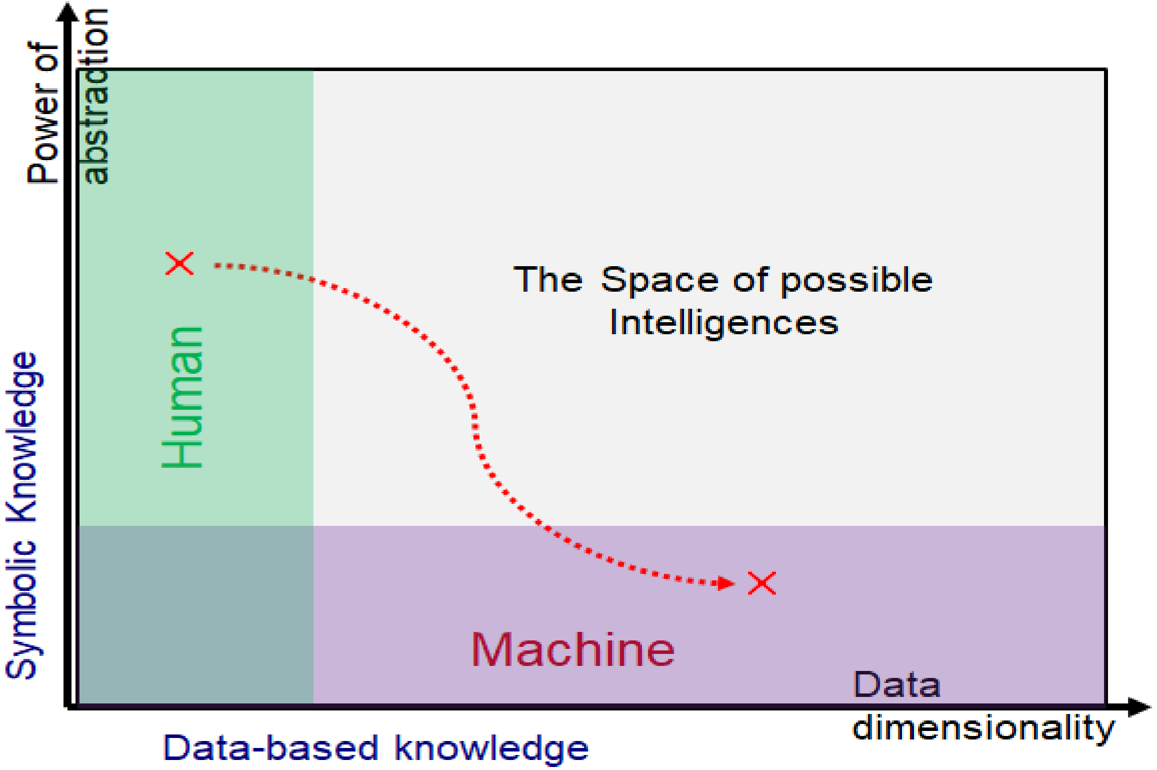
- ❑ Autonomous systems encompass a multi-faceted concept of intelligence.
  - There are multiple intelligences, each characterizing the ability to perform tasks in different contexts;
    To say that “S1 is smarter than S2” is meaningless without specifying the task(s) and the criteria for success.
  - 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*”.

- It is not enough to set out ethical principles: how do we implement them for a given type of application and context by a set of properties, which can be monitored and tested?
  - For autonomous vehicles safety implies collision avoidance and compliance with traffic rules.
  - For ChatGPT safety implies, among other things, not to generate deepfakes.

❑ 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 <u>qualification exams</u> 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 <u>fundamental differences</u> 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.
- Human decision-making implies responsibility for the consequences of actions or omissions.

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

✓ <u>Strive to overcome current limitations with clarity</u>, 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.

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

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? – AI meets Systems Engineering

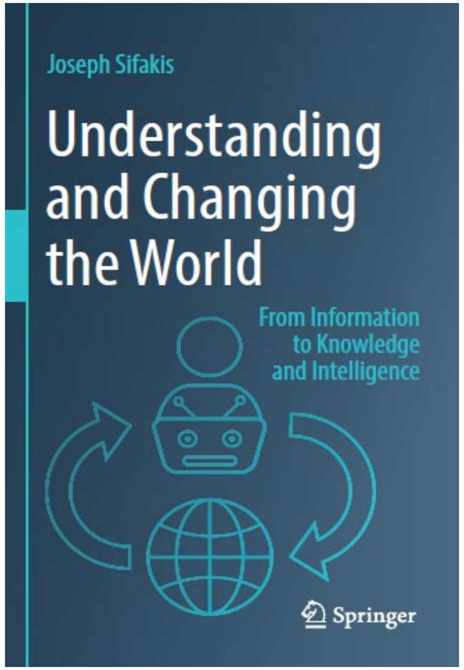
- The development of autonomous systems requires a marriage between ICT and AI, which poses non-trivial technical problems. New trends are disrupting traditional critical systems engineering.

    - 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.


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

Thank you

Joseph Sifakis, Testing System Intelligence, arXiv:2305.11472 [cs.AI]