AI: Today and Tomorrow

Joseph Sifakis

Verimag Laboratory

Digital Humanism Summit on AI and Democratic Sustainability

Vienna

July 3, 2023

About Intelligence – The Concept

- There is currently a great deal of confusion about what intelligence is and how it can be achieved.
  - The spectacular rise of AI, is accompanied by a “frenzy of optimism” 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.
  - A mythology has developed around the issue of the "ultra-intelligence" of machines, according to which they will eventually surpass human intelligence and that we may end up being pets for the machines.
  - Some believe that machine learning and its subsequent developments will enable us to meet the intelligence challenge – this is only a matter of time!

- 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: they outperform humans in games and are capable of performing a wide range of tasks, including some that are obviously sensory sensitive, e.g. AlphaGo, BERT, Chat GPT, DALL-E,..
  - Machines cannot surpass humans 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!

Toward AGI – The Vision

❑The accelerated convergence observed between ICT and AI is leading to *autonomous systems,* a big step from weak AI to AGI.

- Today weak AI gives us the elements to build intelligent systems but we have no principles and techniques to synthesize them e.g. like we build bridges and buildings.

❑ *Autonomous systems* support a paradigm of intelligent systems that goes beyond machine learning systems, which are often specialized transformational systems

- 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.
- 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-trusted AI systems and by difficult systems engineering problems unrelated to agent intelligence.

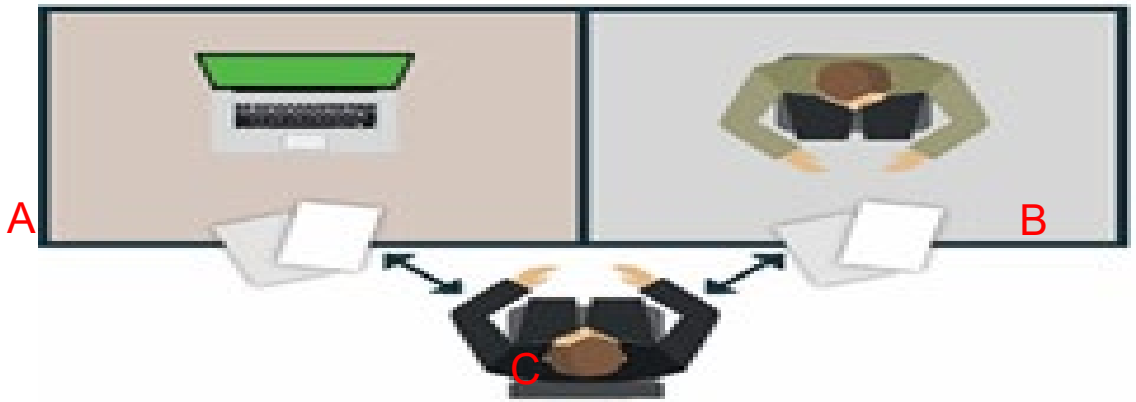
Overview

- Human vs. Machine Intelligence
- Can we Trust AI Systems?
- AI Tomorrow

Human vs. Machine Intelligence – The Turing Test

"Can we experimentally distinguish a computer from a human by analysing their answers to a series of questions?"

- 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 that the Turing test is a special case, where the task is a conversation game.

Human vs. Machine Intelligence – An Interesting Analogy

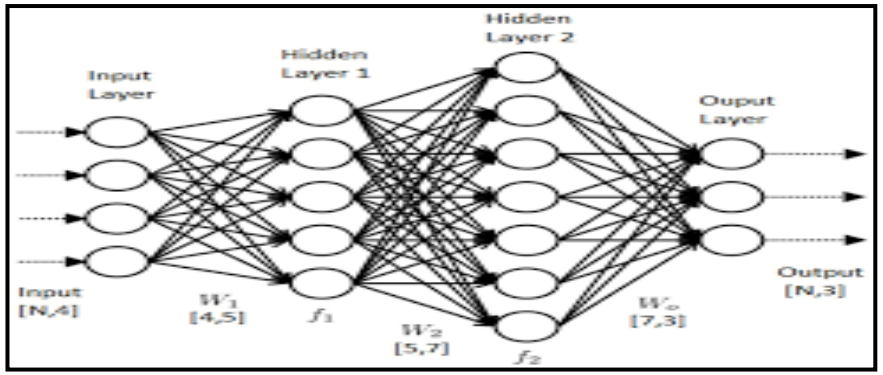
Fast thinking vs. Slow thinking (D. Kahneman's "Thinking Fast and Slow")

System 1: "Fast" Thinking
- Non-conscious – automatic – effortless;
- Without self-awareness or control;
- Handles all kind of empirical implicit knowledge e.g. walking, speaking or playing the piano.

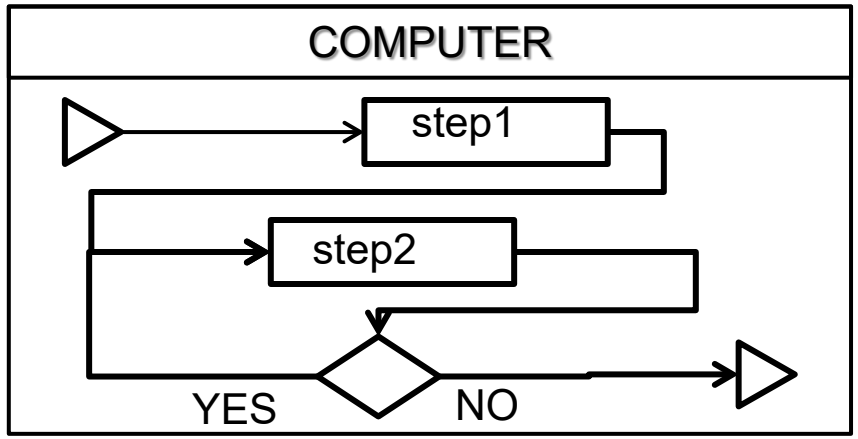
System 2: "Slow" Thinking
- Conscious – controlled– effortful;
- With self-awareness and control;
- Is the source of any reasoned knowledge e.g. mathematical, scientific, technical.

Neural Networks vs. Conventional Computers

NEURAL NETWORK

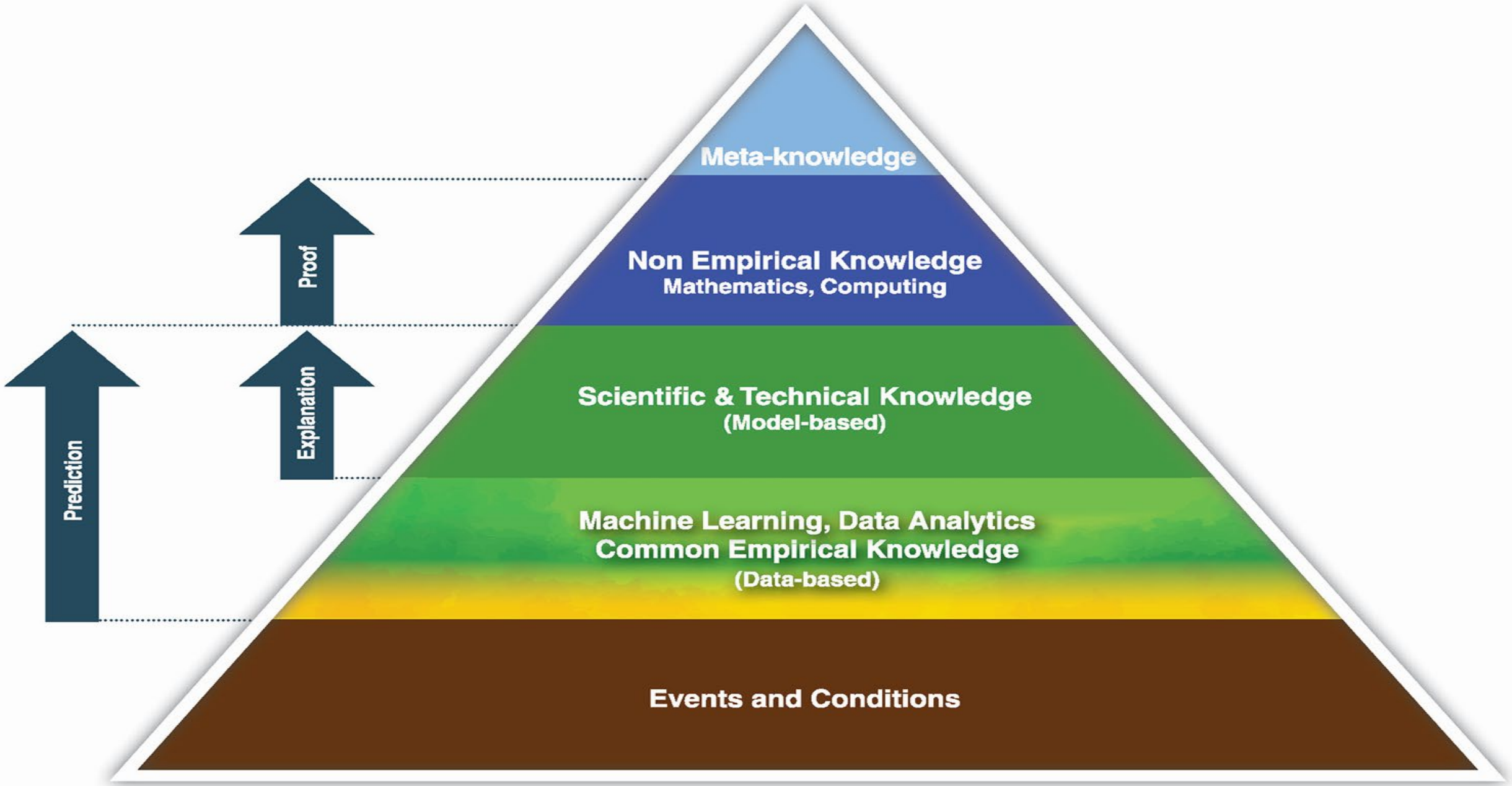


- Generate empirical knowledge after training (Data-based knowledge) – do not execute algorithms
- Distinguish "cats from dogs" exactly as kids do – Cannot be verified!

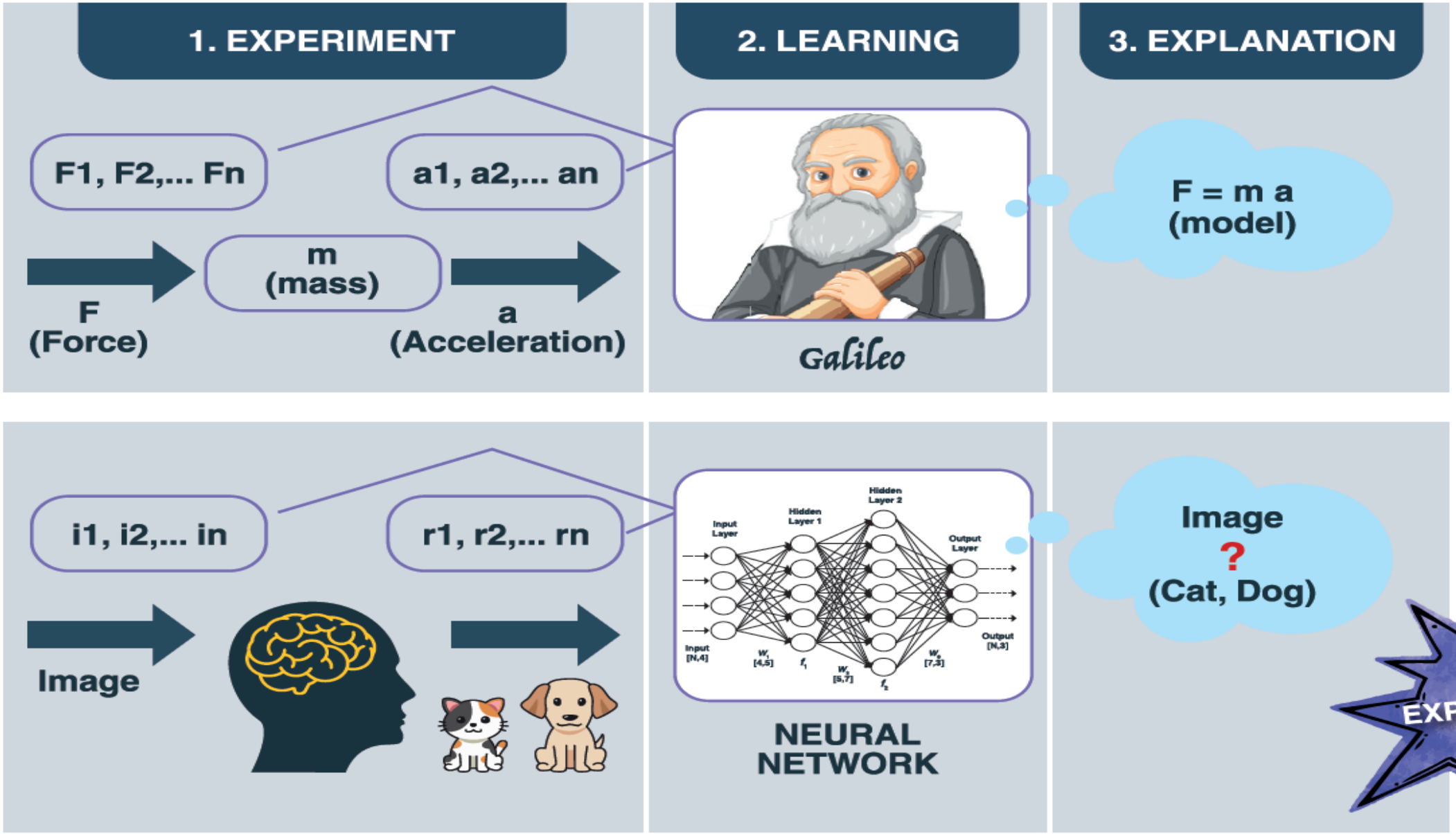


- Execute algorithms (Model-based knowledge).
- Deal with explicitly formalized knowledge
- Can be understood and verified!

Human vs. Machine Intelligence – The Knowledge Hierarchy

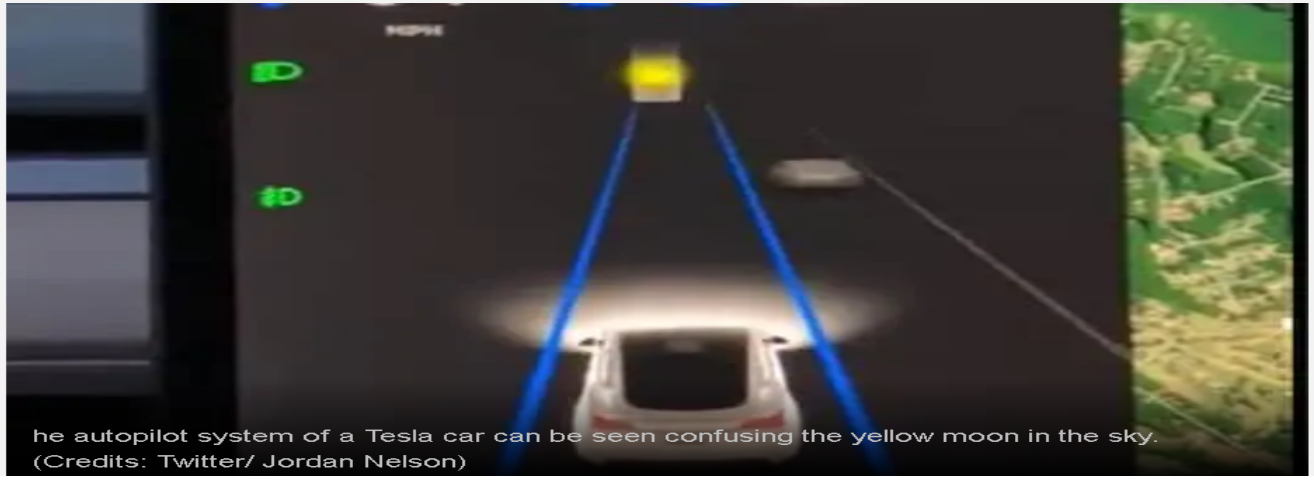


Human vs. Machine Intelligence – Scientific vs. ML-generated Knowledge



Human vs. Machine Intelligence – Common Sense Knowledge

Tesla's Autopilot Feature Mistakes Moon for Yellow Traffic Light, Watch Video



In a viral video shared by Twitter user Jordon Nelson, the autopilot system of a Tesla car can be seen confusing the yellow moon in the sky with a yellow traffic light.

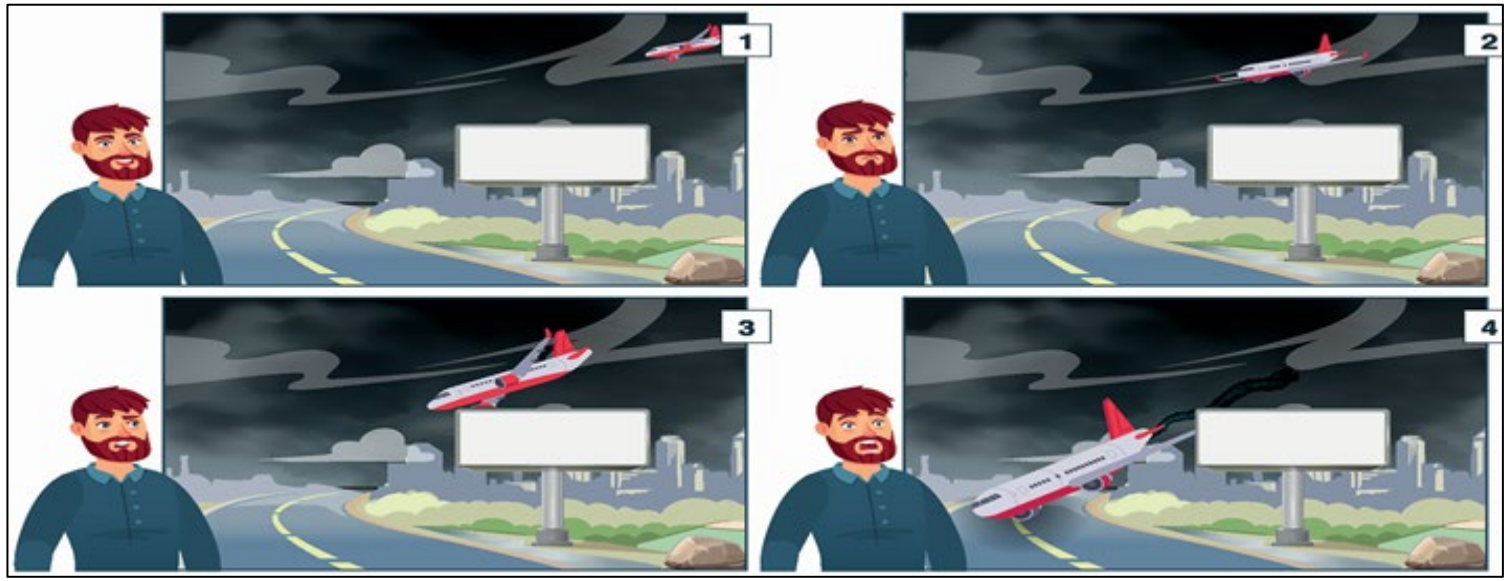
- TRENDING DESK
- LAST UPDATED: JULY 26, 2021, 17:02 IST
- FOLLOW US ON: Facebook Twitter Instagram Telegram Google News

BUZZ STAFF

- To match human-level performance, systems should be able to deal with common sense knowledge.
- Human mind is equipped with a semantic model of the world:
  - a vast network of knowledge progressively built and automatically updated throughout life by learning and reasoning, and involving concepts, cognition rules and patterns;
  - used to interpret sensory information and natural language.
- Human understanding combines:
  - bottom-up reasoning from sensor level to the semantic model of the mind;
  - and top-down reasoning from the semantic model to perception.

Human vs. Machine Intelligence – Common Sense Knowledge

- We recognize a stop sign partially covered with snow because the image matches a conceptual model of a stop sign with its properties (size, color and vertical position).
- In contrast, neural networks must be trained to recognize stop signs in all possible weather conditions.



- This sequence of images is almost instantaneously interpreted as an aircraft accident using common sense knowledge.
- In contrast, a machine although it may be able to recognize the objects in each frame lacks knowledge allowing to infer the same conclusion.

❑ For machines to match human situational awareness, they must be able

- to progressively develop knowledge about their environment, in particular to understand entirely new situations;
- to combine learning and reasoning as well as concrete sensory information with symbolic knowledge.

This is probably the most difficult problem to solve, as shown by the poor progress made so far in the semantic analysis of natural languages.

Overview

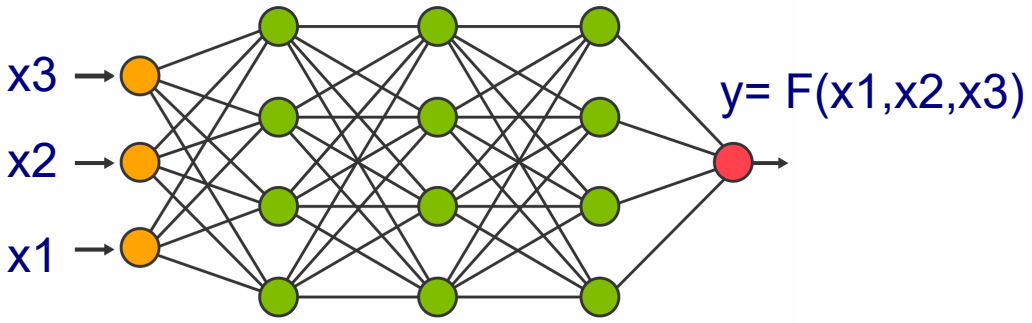
- Human vs. Machine Intelligence
- Can we Trust AI Systems?
- AI Tomorrow

Can we Trust AI Systems? – Explainability of NNs

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

❑ NN explainability : characterize the I/O behavior of a NN from the behavior of its elements

- For feed-forward networks this is theoretically possible as the output can be computed as a function of the inputs given the function computed by each node: *φ(weighted_sum_of_inputs),* where φ is an activation function.
- Approximations of F can be computed for special classes of feed-forward NN with simple activation functions.



❑ Two methods for system property validation: 1) verification (reasoning on a model); 2) testing controlled experiment

Verification methods allow validation of system properties on a behavioral model.

- in particular, it is possible to verify universally quantified properties (safety and security).
- as neural networks cannot be explained, they cannot be verified in the current state of knowledge.

The test methods allow validation of the observed behavior of a system in response to external stimuli.

- properties are subject to observability and controllability constraints;
- properties with universal quantification (safety and security) can only be falsified.

Can we Trust AI Systems? – System Properties

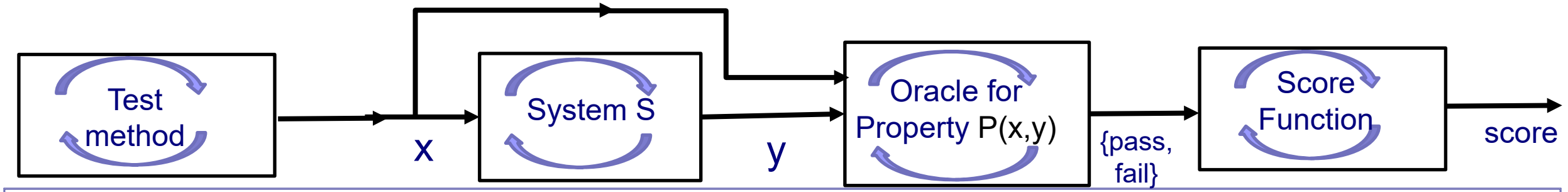
- Systems engineering focuses on three main types of properties:
  - Safety properties meaning that the system, during its execution, will never reach "bad states" characterized by explicit conditions on its variables.
  - Security properties mean that the system is resilient to attacks that threaten data integrity, privacy and system availability.
  - Performance properties characterize technical and economic criteria concerning the resources and their exploitation.

- Epistemic and methodological imperatives demand that assertions that a system satisfies a property be accompanied by its rigorous definition and associated validation method.
  - "*Waymo has now driven 10 billion autonomous miles in simulation*" (July 2019) - This is not a technically defensible argument for the safety of the actual system: simulated miles must be related to "real miles" to show that the simulation deals fairly with the many different situations, e.g., different road types, traffic conditions, weather conditions, etc.
  - "*Responsible AI*" implies that the development and use of AI meets criteria such as fairness, reliability, safety, privacy and security, inclusiveness, transparency, and accountability, which are difficult, if not impossible, to assess.
  - "*AI alignment*" rings hollow; what it means to align a conversational agent with human values as we do not even understand how the human volition emerges and the associated value-based decision-making system works.

Can we Trust AI Systems? – Validation by testing

- Testing allows providing experimental evidence that a system y=S(x) satisfies a property P(x,y) using a framework:
  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.*

" S satisfies P" means that for any possible input x of S and corresponding y, the property P(x,y) is satisfied.



- Test method: How to choose among the possible test cases and decide whether the process is successful or not?

  1. Efficiency Function: *efficiency* such that *efficiency*(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* such that *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:

efficiency(X1)=*efficiency*(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 unambiguous criteria.	Statistical evidence? / Statistical reproducibility?
Simulated Self-driving systems	Formally specified properties e.g. Traffic rules	Test method for driving scenarios?	Automated Analysis of system runs	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 hampered by adversarial examples -- observationally equivalent test cases give different scores;
  - is limited to properties that can
    - be rigorously specified, which excludes Q/A relations for natural language transformers;
    - be observed, which excludes "human-centric" properties e.g., intentionality, belief, awareness.

Overview

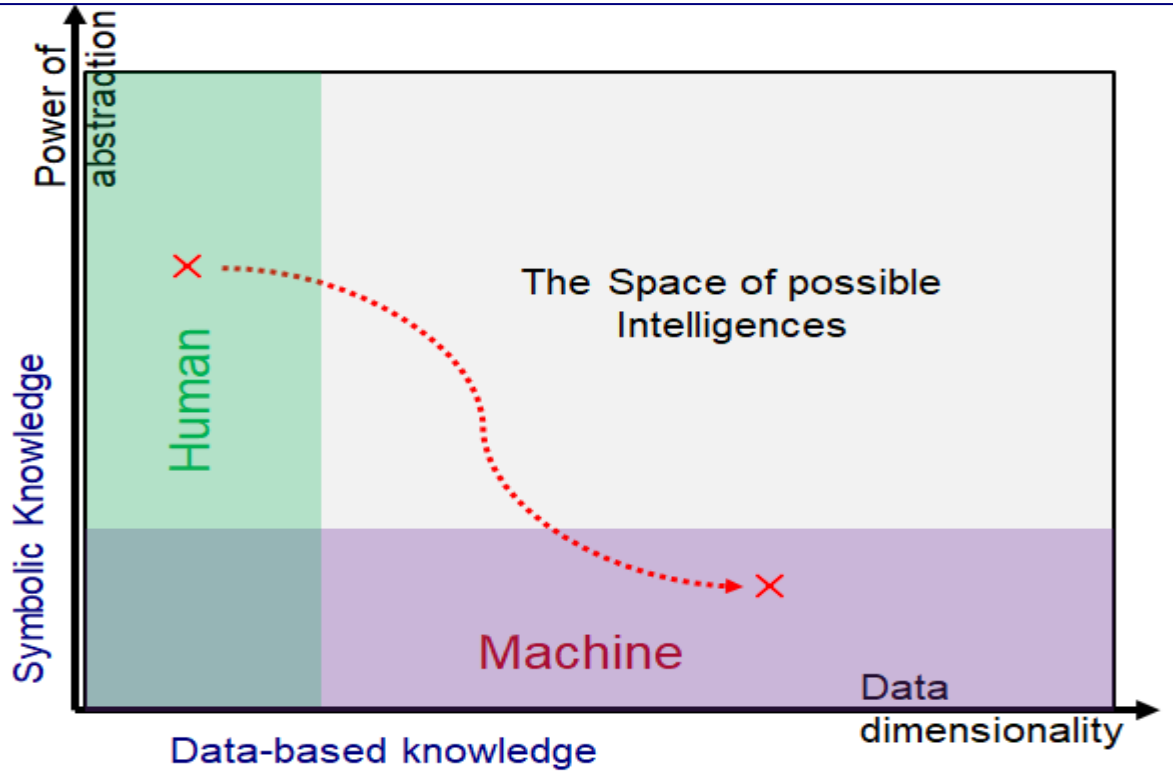
- Human vs. Machine Intelligence
- Can we Trust AI Systems?
- AI Tomorrow

AI Tomorrow – The Space of Possible Intelligences

- The replacement test relativizes and generalizes the concept of intelligence suggesting that intelligence is not one-dimensional
  - To say that "S1 is smarter than S2" is meaningless without specifying the task(s) and the criteria for success. There are multiple intelligences, each characterizing the ability to perform tasks in different contexts;
  - Human intelligence is not "general purpose"; it is the result of historical evolution in a given physical environment. If human intelligence is the benchmark, AGI 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



AI Tomorrow – Can We Trust AI Systems

- Tendency to ignore the limitations of validating intelligent systems according to established systems engineering criteria:
  - LLMs' success on various benchmarks designed to model meaning-sensitive tasks, is considered sufficient evidence that they understand natural language. However,
    - The training data for LLMs do not account for meaning;
    - Without adequate coverage criteria, it is impossible to demonstrate that the benchmarks are not free of bias.
  - Many works superficially attribute mental attitudes such as belief, desire and intention to autonomous systems: "*we cannot show that an agent always does the right thing, but only that its actions are taken for the right reasons*". However,
    - Testing ethical criteria means that machines "understand" the impact of their choices and can emulate value-based decision mechanisms currently poorly understood.
    - Moreover, the "Chinese room argument" shows that the ability to understand cannot be discerned by black box tests.

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

- ✓ Admit the need for rigorous validation methods, and not get lost in a bogus and irrational debate about the human-centric properties of machines.
- ✓ Strive to overcome current limitations with clarity, developing new foundations, and possibly revising epistemic and methodological requirements where necessary.

AI Tomorrow – A New Kind of Science

- We can leverage the complementarity between humans and machines to accelerate the development of knowledge.
  - Humans are limited by *cognitive complexity* in extracting knowledge from high-dimensional data, whereas they can develop symbolic knowledge using powerful abstraction mechanisms, e.g. induction, metaphors, analogies, creative thinking.
    ⟹ Useful scientific theories involve a small number of independent concepts and parameters. Often to study complex phenomena e.g. economic, we do simplifications that may prove to be unrealistic.
  - Machines are capable of computational intelligence by creating knowledge from high-dimensional data, but their current ability to create and apply symbolic knowledge is limited.

- We can generate a new type of knowledge, between scientific knowledge and implicit empirical knowledge, allowing predictability and only limited understanding.
  - Thanks to supercomputers and AI, we can build *neural oracles* allowing predictability of complex phenomena e.g. geophysical, economic, social, etc.
  - If a neural oracle is explainable, then its models can serve as a basis for the construction of a "theory" explicating the relations between the observables.
  - The knowledge so produced is entirely discovered by machines and impossible to apprehend conceptually.
  - Using this kind of knowledge, especially to make critical decisions, should give us pause.

- Knowledge production/application is no longer the preserve of humans – two possible scenarios:
  - **Symbiosis**: human-centred machine-assisted development of science and technology
  - **Divergence**: parallel development of a "para-science “

AI Tomorrow – Social Impact

Social acceptance depends on the role of institutions that largely contribute to shaping the public opinion, e.g. about what is true, right, safe or secure.

- The trustworthiness of infrastructure and 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!

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

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

  - I do not believe that computers can surpass the intelligence of their creators.
    But it would be possible for creators to be enslaved to computers because they would be overtaken by the complexity of their management or because of laziness and "for convenience" .
    And that would be a catastrophic scenario for humanity!

Thank you

Joseph Sifakis “Testing System Intelligence” https://arxiv.org/abs/2305.11472