CURRENT AND FUTURE CHALLENGES IN SYSTEMS ENGINEERING OF INTELLIGENT SYSTEMS

FROM HYPE TO PRACTICE USING EXAMPLES FROM MEDICAL DEVICE DEVELOPMENT

Michael Kremliovsky, INCOSE San Diego Mini-conference, Dec 3, 2022

WHAT THIS PRESENTATION IS ABOUT

- An engineering framework which gives Systems Engineers a way of thinking about “AI”
- Evolution of Cyber-Physical Systems to Intelligent Systems
- Key regulatory and ethical challenges, now and in the (near) future

a systems engineer trying to understand what to do with artificial intelligence

… DALL-E MAY ALSO DO A GOOD JOB

Still life photorealistic oil painting with vegetables, cabbage, pepper, cucumber, grapes, and a glass of red wine in the light of sunset on the table with flowers

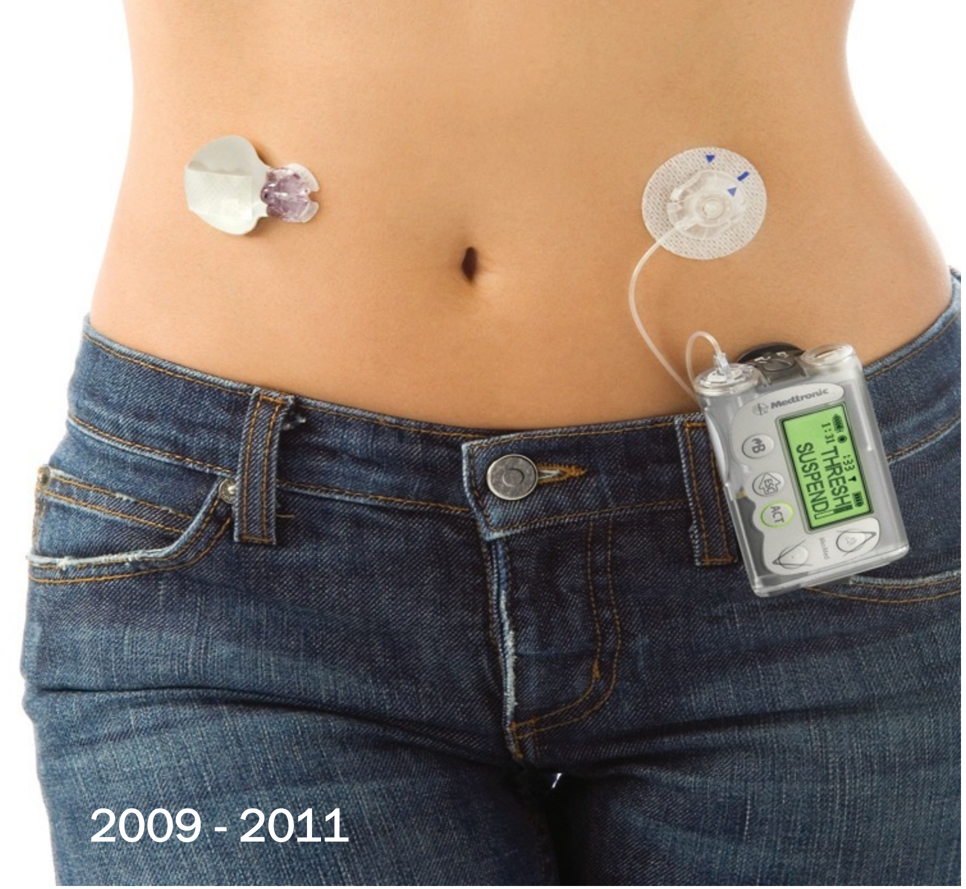
AGENDA

- On Terms and Definitions
- Intelligent Agents and Intelligent Systems
- Regulatory Challenges
- Ethically Aligned Design

TERMS & DEFINITIONS: CYBER-PHYSICAL SYSTEMS

- Why defining terms is important?
  - Linguistic Relativity a.k.a. Sapir-Whorf hypothesis (or Whorfianism): language affects world views and cognition
  - Language of Thought (LOTH): we think in linguistic concepts (tokens, semiotics, semantics)
  - Communication: we need to understand each other
    (bad example: “Artificial Intelligence”; good example: “Artificial Neural Network”; even better: “Convolutional Neural Network”)

- **Cyber-Physical System (CPS):** “Devices that incorporate a mechanism that is controlled or monitored by software and electronic components and that is tightly integrated with the computer networks and its users; such system can exhibit multiple, distinct behavioral modalities that may change with context.” (US/NIST Special Publication 1500-201)
- IEC 60601 group of standards, synonymous: Programmable Electrical Medical System – PEMS.



MEDICAL DEVICES AS CYBER-PHYSICAL SYSTEMS

EVOLUTION OF CYBER-PHYSICAL SYSTEMS

- Embedded Software Controls of Electro-Mechanical (“Physical”) System
- Interface to Human Operator and Data Display
- Network Integration for Data Communication and Maintenance
- System-of-Systems Integration: several systems working in cooperation

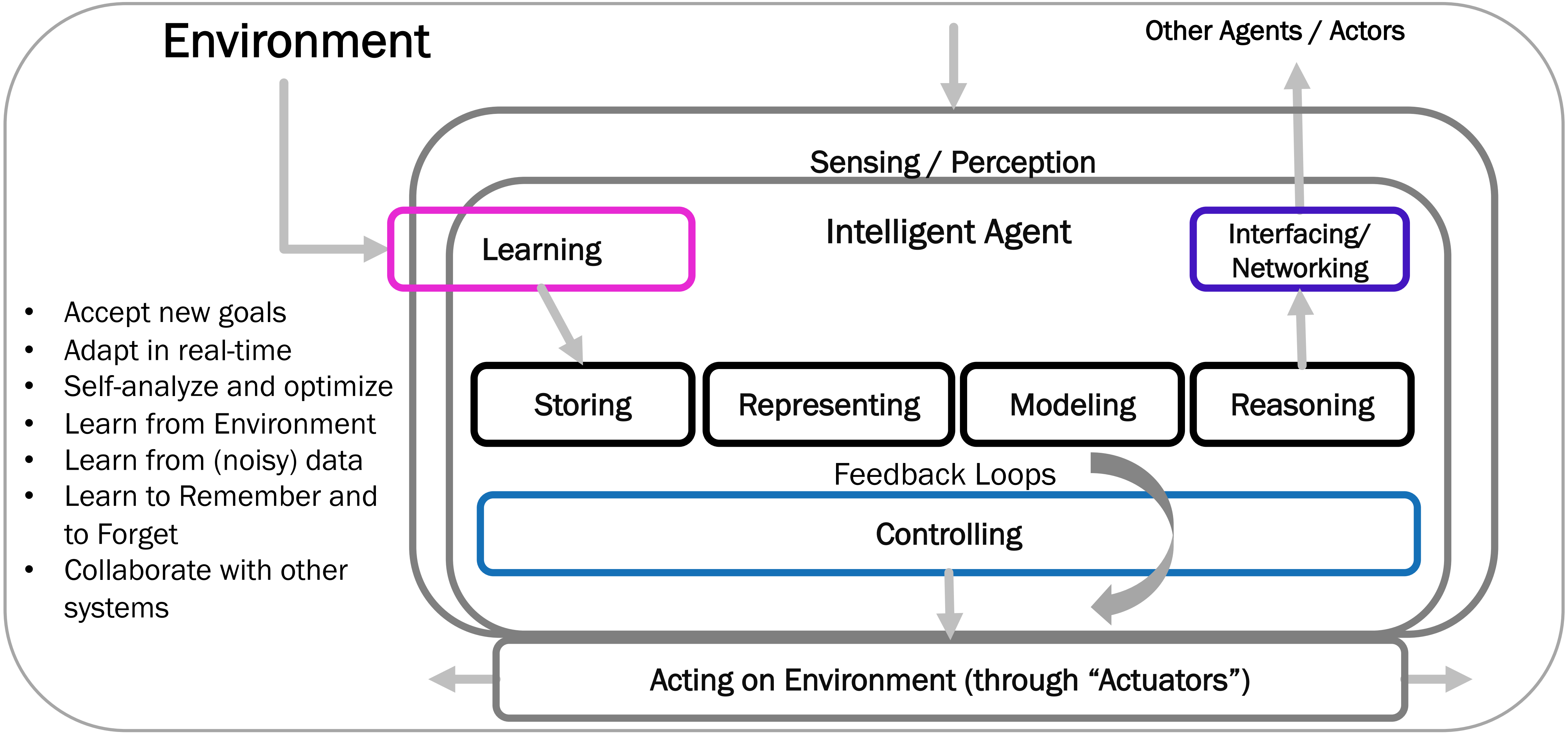
- Situational Awareness and Decision Autonomy

INTELLIGENT AGENTS* ARE CYBER-PHYSICAL SYSTEMS

- Agent – something that acts (*agere,* Latin: to do, to drive)
- An Intelligent Agent (IA) is an *autonomous* entity that directs its activities toward achieving specific goals by making observations of its environment through sensors, processing the inputs, and acting on the environment through its actuators (or effectors).
  - Agent interacts with its environment
  - Agent has externally or internally defined goals
  - Agent receives information about its environment through sensors
  - Agent can process the incoming information to make "decisions" concerning how to act
  - Agent can act toward achieving its goals using its effectors
- Intelligent Agents are not *Automatons* (control mechanism designed to automatically follow a predetermined sequence of operations, or respond to predetermined instructions)

* For the purpose of this presentation, we consider *non-biological* Intelligent Agents or *machines*; see Russell, SJ; Norvig, P (2016), *Artificial Intelligence: A Modern Approach*, 3rd ed., Pearson Education Limited.

INTELLIGENT AGENT: GENERIC ARCHITECTURE



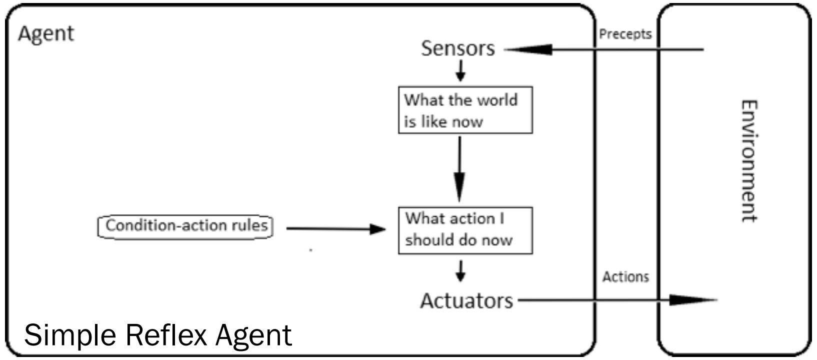
- Accept new goals
- Adapt in real-time
- Self-analyze and optimize
- Learn from Environment
- Learn from (noisy) data
- Learn to Remember and to Forget
- Collaborate with other systems

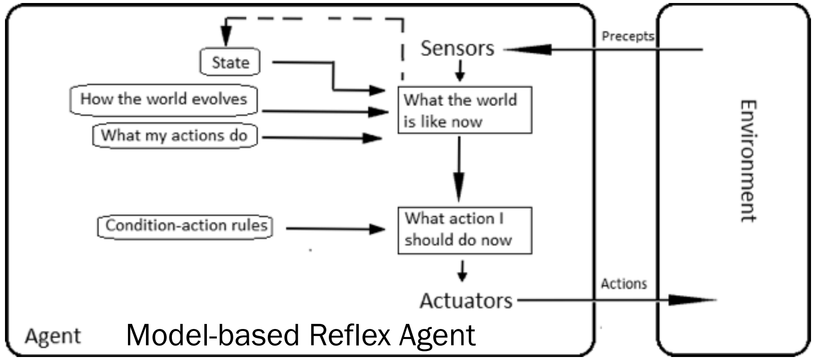
ASHBY'S LAW OF REQUISITE VARIETY (OR 1ST LAW OF CYBERNETICS)

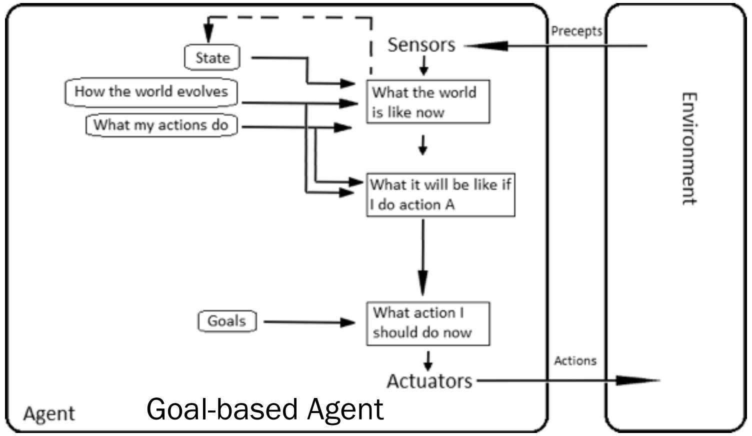
- **Norbert Wiener** defined **Cybernetics** as "the science of control and communications in the animal and machine"
- **Andrei Kolmogorov:** "science concerned with the study of systems of any nature which are capable of receiving, storing and processing information to use it for control"
- **Law of Requisite Variety:** if a system is to be stable, the number of states of its control mechanism must be greater than or equal to the number of states in the system being controlled.
- **Good Regulator Theorem** (Roger C. Conant and W. Ross Ashby): "Every good regulator of a system must be a model of that system" (Int. J. Systems Sci., 1970, vol 1, No 2, pp. 89–97)

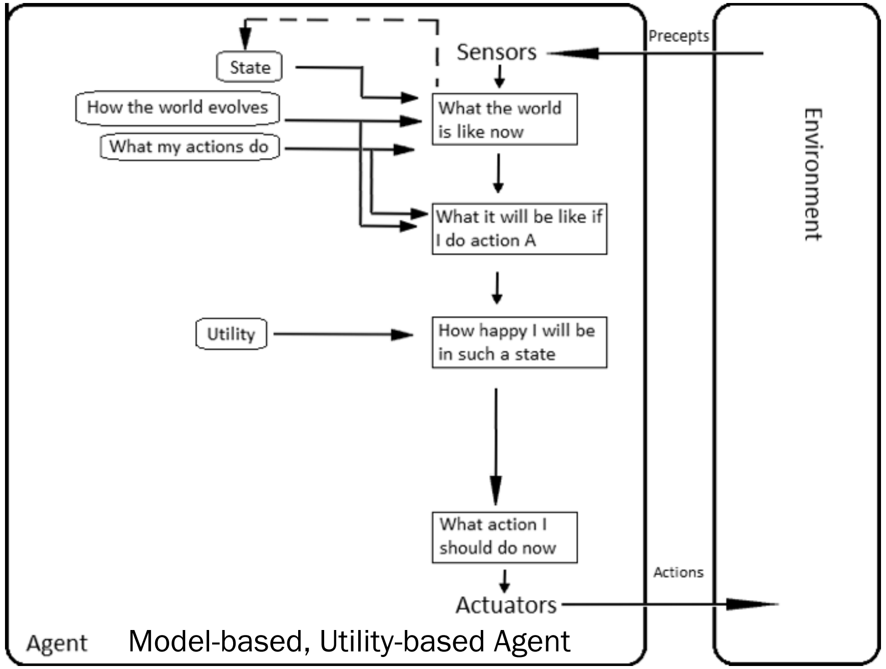
- The Big Elephant in the room of Validation of Intelligent Agents...

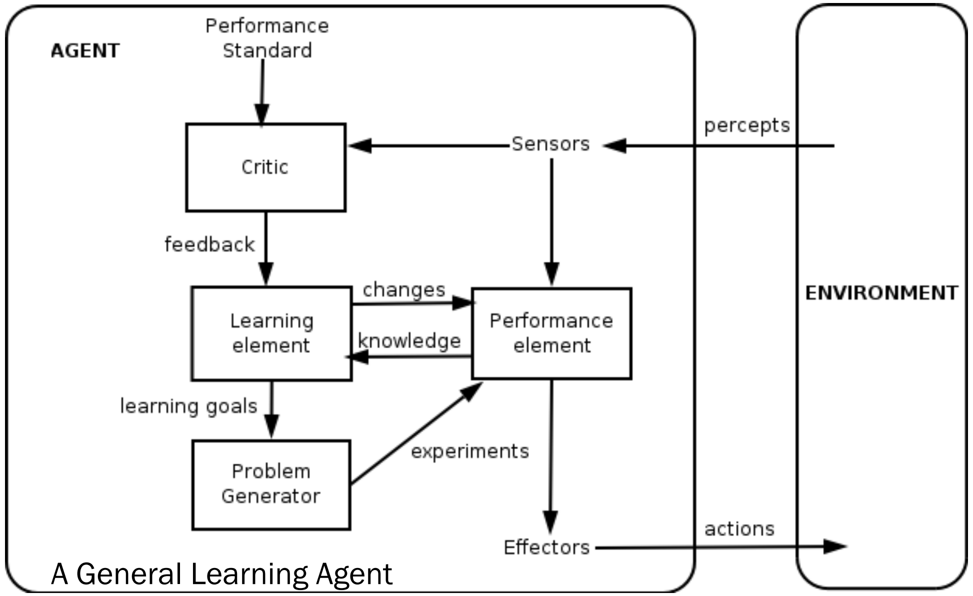
INTELLIGENT AGENTS: SPECIALIZED ARCHITECTURES



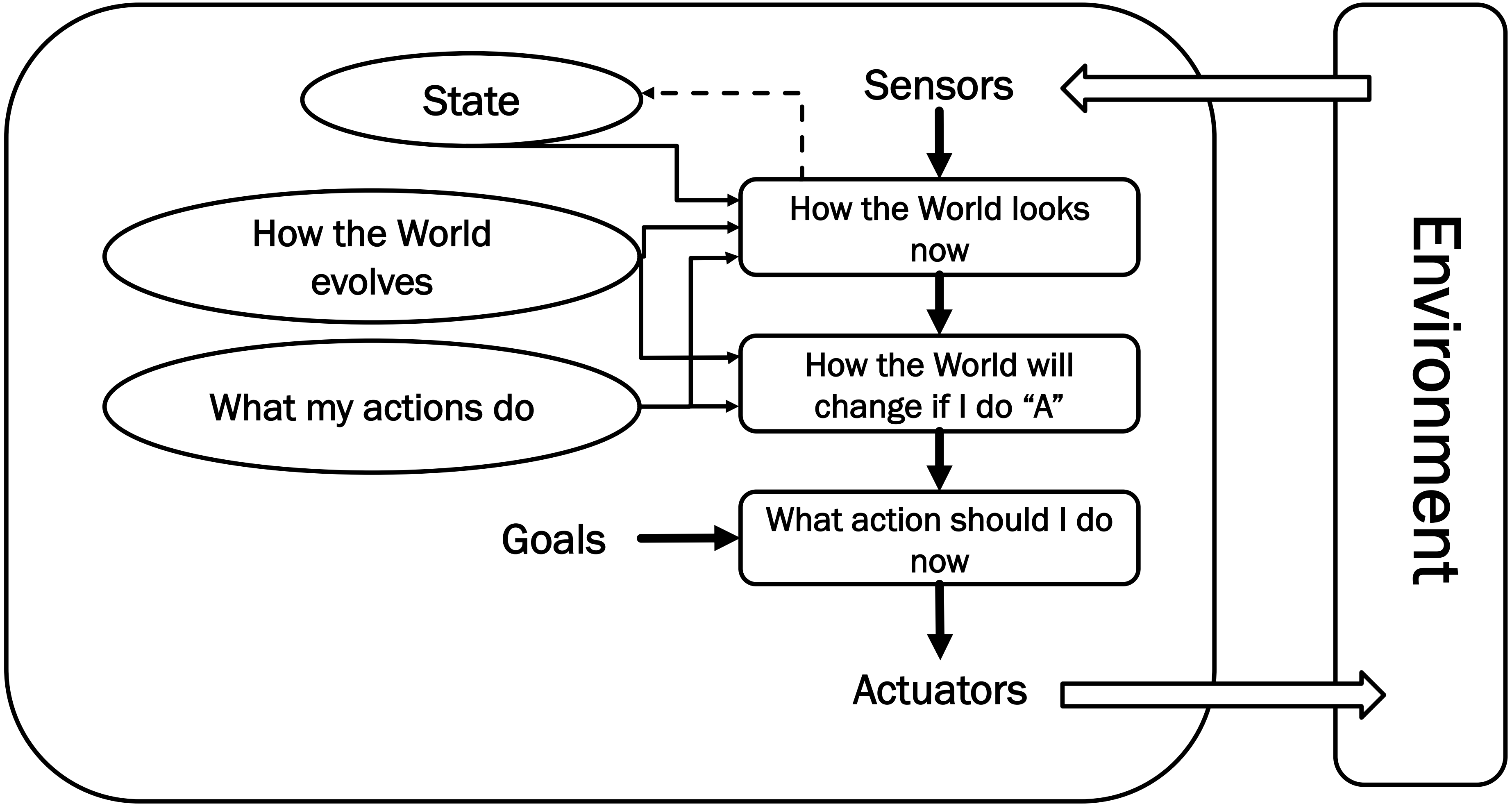








GOAL-BASED AGENT



COMPLICATIONS: INTELLIGENT SYSTEMS

- Extensive Networking and Communication
- Virtualization of storage, computing, and networking infrastructure
- Distributed Agents: different components are loosely connected / integrated and may come from different manufacturers
- Cooperating Agents: several agents sharing goals and working in cooperation (e.g. swarms)
- Software Agents: platform-agnostic software systems using communication channels to affect environment (e.g. chatbots via social interaction)

- **Generalization of Agents: Intelligent Systems**

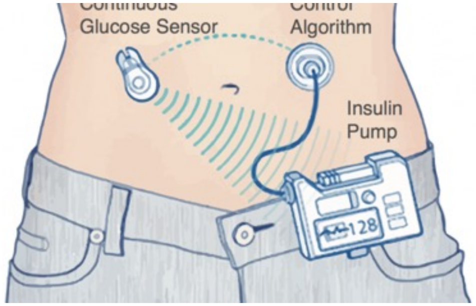
AUTONOMY OF INTELLIGENT SYSTEMS

- *No Autonomy*. Based on inputs, the System suggests a decision (or multiple ranked decisions). Human retains full responsibility for accepting or rejecting the suggestions and executing subsequent actions.
- *Supervised Autonomy*. The System operates under constant run-time human supervision with an option of human override or returning the control to a human.
- *Unsupervised Autonomy*. The System is designed to operate autonomously at all times except normal maintenance and/or goal-setting episodes requiring human intervention.
- *Full Autonomy.* The Intelligent System sets its own goals and operates without any intervention or supervision of humans.

- A note on "Artificial General Intelligence" ("… the intelligence of a machine that can understand or learn any intellectual task that a human being can…"): introduction of this term is a result of not understanding or not willing to define what Intelligence is in the first place.

INTELLIGENT SYSTEMS IN HEALTHCARE

- Systems with <u>no</u> Decision Autonomy are not in the scope here.
- Next-Gen Clinical Decision Support (CDS) Systems
  - Professional (diagnostic systems, therapy planning and optimization, predictive biomarking, etc.)
  - Consumer (symptom checkers, self-triage systems, disease management, companion diagnostics, etc.)
- Advanced Autonomous Closed-Loop Systems
  - Personal Therapeutic
  - Surgical – access, precision, speed, efficiency
- Robo-Nurses, Robo-Technicians, Personal Assistants, and Care Guardians



Autonomous Wheelchair by Singapore-MIT Alliance for Research and Technology

RIBA (Robot for Interactive Body Assistance) build by Japan's Institute of Phys. and Chem. Research & Tokai Rubber Industries

Clinical Workflow Assistant

Moxi by Diligent Robotics in Texas Health Dallas trials

DP14 Hawk – Medical Evacuation Autonomous Transport

INTELLIGENT SYSTEMS IN HEALTHCARE: CRAWLING... BUT MIGHT BE RUNNING SOON

THE QUESTION IS NOT WHETHER WE CAN CREATE (VERY) INTELLIGENT SYSTEMS FOR HEALTHCARE BUT RATHER: DO WE NEED AND WANT TO?

INTELLIGENT SYSTEMS AND REGULATIONS

- A lot is covered by domain-specific regulations: telecommunications, automotive, aero-space, **medical device.**
- BUT, just a few examples where there is a deficit of guidance:
  - How to incorporate Machine Learning (ML) into the IS programming?
  - How to create/design a human-machine interface supporting human development?
  - How to validate intelligent systems in real-world environments?
  - How to ensure appropriateness of data collection and protection of privacy?
    - **The need for absolute privacy* comes here!**
  - How to ensure fair (and legal!) marketing practices of intelligent systems?
  - How to set effective the market/public feedback mechanisms?

* By *absolute privacy* we mean the data security design when the data is not accessible by the third parties, under any circumstances.

WHY REGULATIONS ARE IMMINENT

- Three ways to establish a healthy business environment:
  - Self-governance: every party (public and business) wins by cooperating and following self-imposed guidelines
  - Standards: every party wins by conforming to standard specifications and processes
  - Laws: a unified set of rules and the corresponding enforcement framework levels the field in public interests
- Use of Intelligent Systems in business processes and products provides significant competitive advantage – no incentive for business to collaborate on shared development strategies (open sourcing is fine)
- Replacing humans with Intelligent Systems creates lucrative opportunities for business automation (as naively perceived, but often resulting in unethical activities or even harm)
- The existing infrastructure is not ready for autonomous systems, it is not friendly to robots
- Recently published guidance documents regarding IS development (country-specific and by professional organizations) contain a long wish list of attributes without hints on how to achieve the stated goals
- Human-machine interfaces require a very different privacy protection model: *process-and-forget*

A FEW DEFINITIONS, AGAIN

- **Understandability** (or equivalently, intelligibility) denotes the characteristic of a model to make a human understand its function – how the model works – without any need for explaining its internal structure or the algorithmic means by which the model processes data internally.
- **Comprehensibility** (in the context of ML model) refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion.
- **Interpretability** is defined as the ability to explain or to provide the meaning in understandable terms to a human.
- **Explainability** is associated with an interface between humans and a decision-making machine that is both an accurate proxy of the underlying model and comprehensible to humans.
- **Transparency** of a system is *understandability* of the underlying models plus disclosure of information on the system's origin, legality and accountability of its operation.

A FEW DEFINITIONS: BELIEVES VS. ENGINEERING

- Fairness*
  - Unobservable theoretical construct (cannot be measured directly, must be inferred)
  - Different theoretical understandings possible depending on the context – fairness is a *contested construct*
  - Operationalization: individual vs. group-level fairness, parity (every group has the same outcome) vs. calibration (same parameters lead to the same outcome across groups)
  - In general, it is impossible to build a fair system without specifying the context and the measurement criteria; however, then the system is only going to be fair within *THAT* context (*content validity*)
  - Operationalization of fairness lacking a due process of disputing (justice) lacks *consequential validity*
- Bias (in Machine Learning)** – "… results that are systematically prejudiced due to erroneous assumptions…"
  - Bias, of one sort or the other, is present in *any* model, bias is impossible to avoid in practical applications
  - A known bias can (and should) be disclosed; an unknown bias is a problem (ground truth is often unavailable)
  - Machine Learning only is as good as its training data (incomplete data = algorithmic bias)
  - Sample-size disparity (equalizing statistical assumptions from unequal samples), reward hacking (incorrect reward function), …

* For a review of the topic, see: arXiv:1912.05511v1 (Abigail Z. Jacobs and Hanna Wallach "Measurement and Fairness", working paper)

** Mireille Hildebrandt "Machine Learning and Society: Impact, Trust, Transparency", MIT Press 2020 (forthcoming)

“ETHICAL DILEMMAS” (AN EXAMPLE)

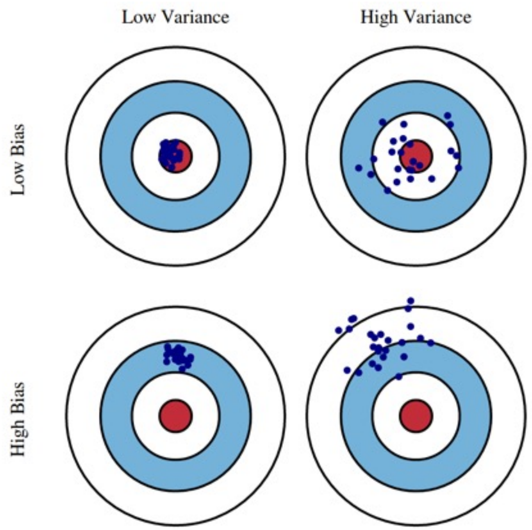
- What should the Autonomous Vehicle Do?
  - Option 1: stay in lane, endanger the Baby
  - Option 2: deviate, endanger the old Lady

WRONG ANSWER!

- THE CORRECT ANSWER:
  <u>bring the Vehicle to the safest state (rest) in the shortest time without violating the traffic rules</u>

- Machines are *amoral* (lack morality): cannot reason on moral values and/or hypothetical circumstances

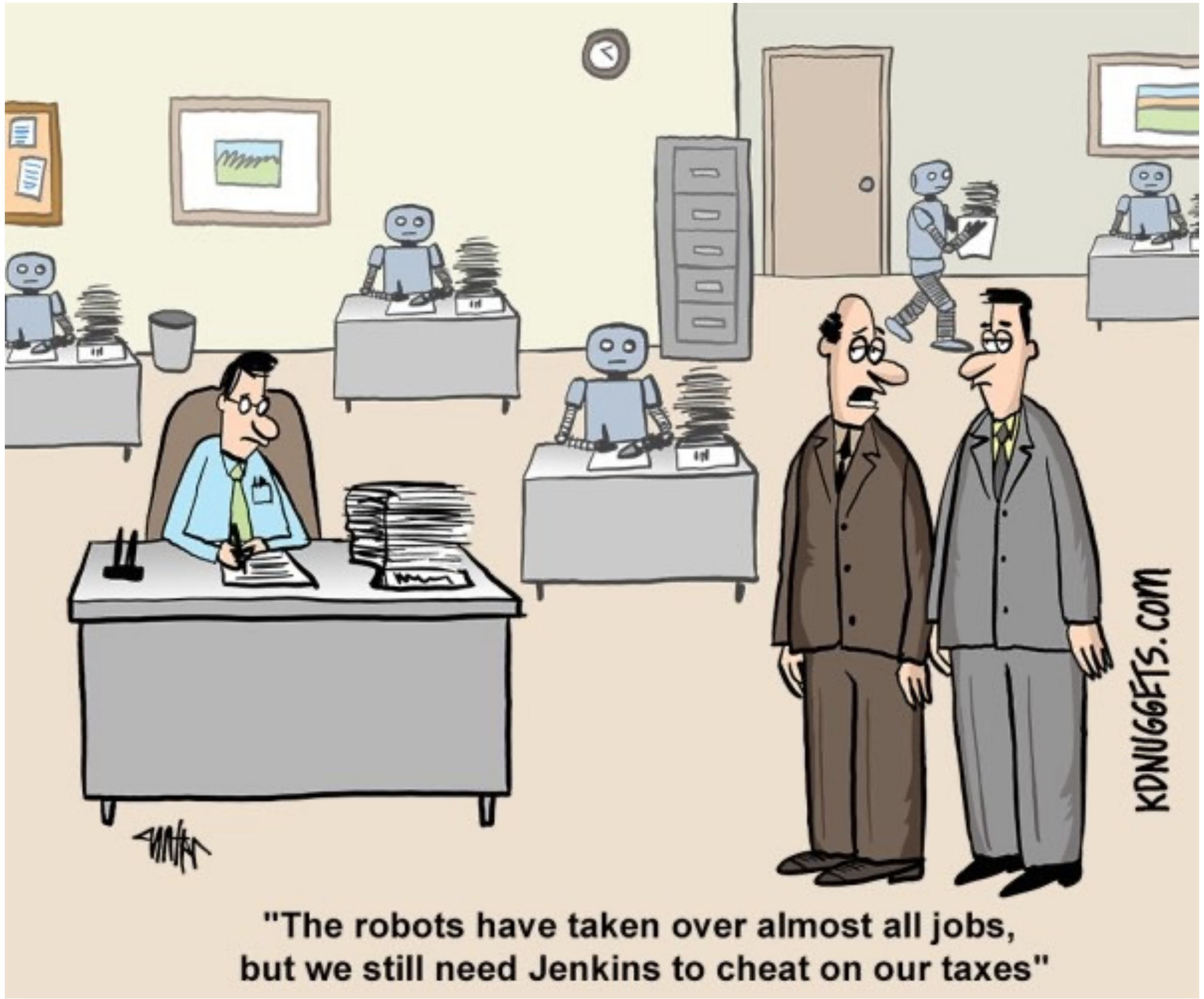
DESIGN TRADEOFFS: ONE CANNOT HAVE IT ALL…



- Resources-Schedule-Features (can have 2 out of 3)
- CAP (a.k.a. Brewer's) Theorem– it is impossible for a distributed data store to simultaneously provide more than two out of the following three guarantees:
  - Consistency: Every read receives the most recent write or an error
  - Availability: Every request receives a (non-error) response, without the guarantee that it contains the most recent write
  - Partition tolerance: The system continues to operate despite an arbitrary number of messages being dropped (or delayed) by the network between nodes
- Bias-Variance Dilemma – underfitting and overfitting problem
- No Free Lunch (NFL) Theorem: if an algorithm does particularly well on average for one class of problems, then it must do worse on average over the remaining problems (David Wolpert & William Macready, 1997)
- Security/Privacy-Interoperability/Ease-of-Use

CONCLUDING REMARKS

- Intelligent Systems is an important (if not the most important) branch of future Systems Engineering
- We do not have to be stupid (or political) when we design and task Intelligent Systems
- Current regulations are not sufficient for creating a safe marketplace
- Intelligent Systems cannot and should not make human decisions



IEEE GLOBAL INITIATIVE ON ETHICALLY ALIGNED DESIGN FOR AUTONOMOUS AND INTELLIGENT SYSTEMS (A/IS)

- The Institute of Electrical and Electronics Engineers (IEEE): 420,000 members in 160 countries
- Eleven IEEE P7000™ Standards Working Groups
  - IEEE P7000™ - Model Process for Addressing Ethical Concerns During System Design
  - IEEE P7001™ - Transparency of Autonomous Systems
  - IEEE P7002™ - Data Privacy Process
  - IEEE P7003™ - Algorithmic Bias Considerations
  - IEEE P7009™ - Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
- Goals: Human Rights, Well-being, Accountability, Transparency, Awareness of Misuse
- Objectives (measures of the progress):
  - Personal Data Rights and Individual Access Control
  - Well-being Promoted by Economic Effects
  - Legal Frameworks for Accountability
  - Transparency and Individual Rights
  - Policies for Education and Awareness