

Lecture 1

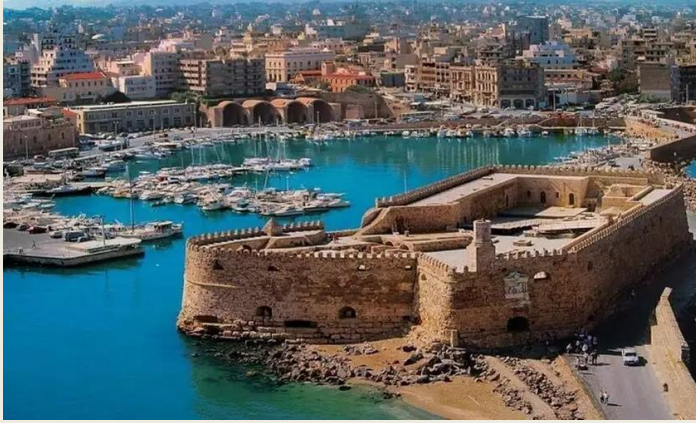
Introduction

Michail G. Lagoudakis

**Intelligent Systems Laboratory
School of Electrical and Computer Engineering
Technical University of Crete
Chania, Crete, Greece**



A few words about me (in pictures)



More words about me

- Research interests
 - Artificial Intelligence (AI)
 - Machine Learning (ML)
 - Decision Making under Uncertainty
 - Reinforcement Learning (RL)
 - Autonomous Robotics
 - Multi-Agent Systems
- Robot Team “Kouretes”



A few words about the course

Course objectives

1. To introduce the students to the basic concepts of Artificial Intelligence (AI) and Machine Learning (ML).
2. To present to the students state-of-the-art applications of AI and ML in medicine and healthcare, focusing on decision-support systems for disease diagnosis and treatment.
3. To develop the students' interest for such systems and for AI and ML in general.

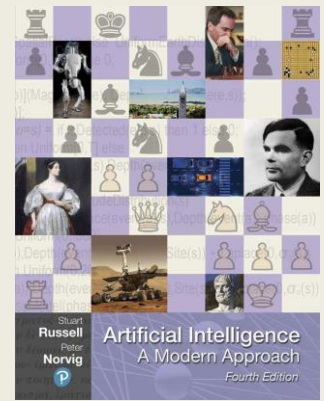


Course Content (7 lectures)

1. **Introduction to AI.** What is AI. History of AI. Symbolic AI and Expert Systems. Statistical AI. Data-driven AI. The Intelligent Agent paradigm. AI for medicine: an overview.
2. **Intelligent Search methods.** Informed Search and Uninformed Search.
3. **Constraint Satisfaction Problems.** Applications in the Medical Domain.
4. **Logic:** propositional logic and first-order logic. Logical Inference.
5. **Decision-making under uncertainty:** Bayesian reasoning, Markov Decision Processes.
6. **Expert Systems** for Medical Decision Support and Assistive Technologies (symbolic AI-based expert systems, Bayesian systems, hybrid approaches).
7. **Machine Learning:** Neural Networks, Deep Learning, and their applications in medicine and healthcare.



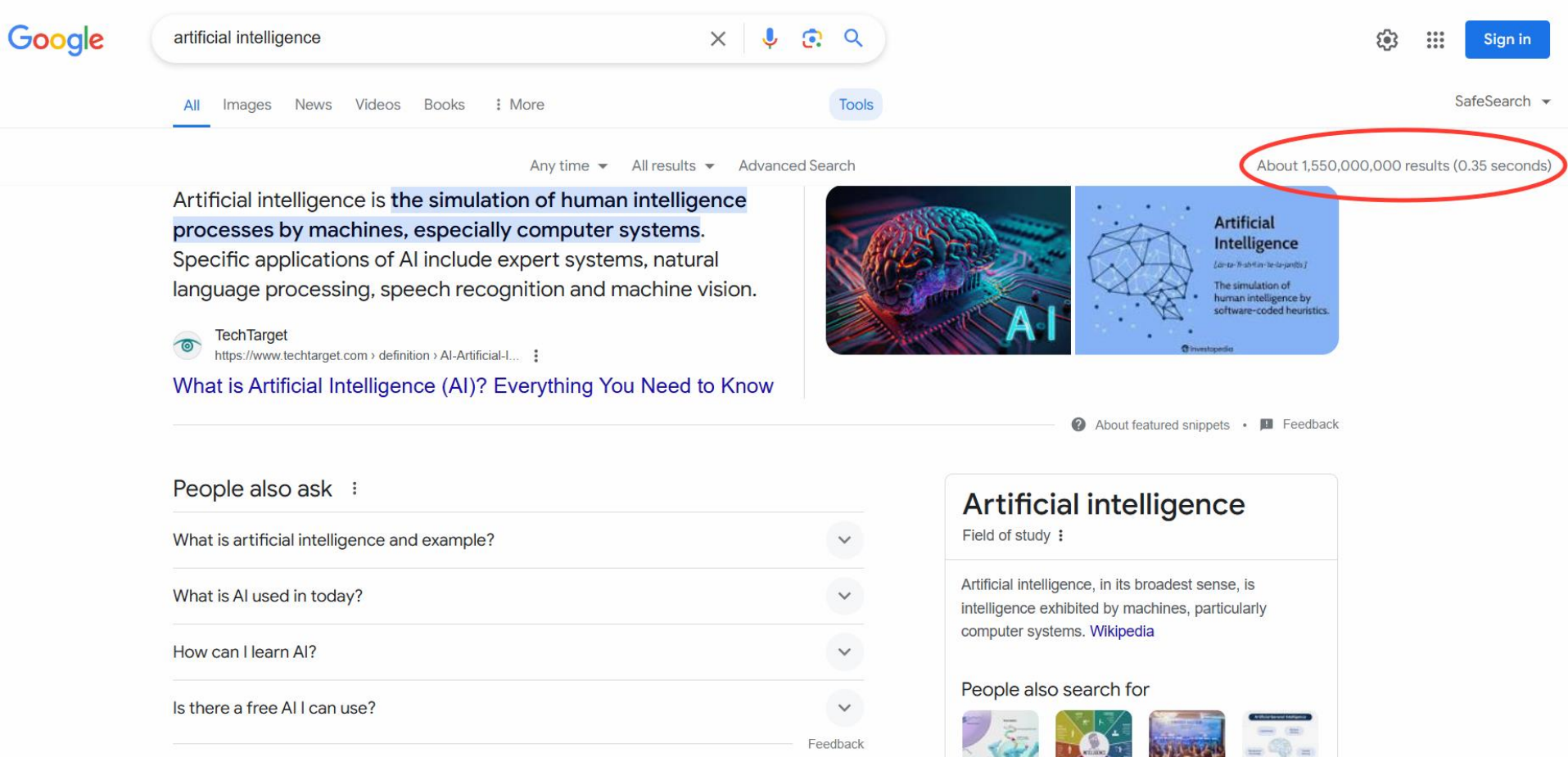
Bibliography



- *Artificial Intelligence: A Modern Approach.* Stuart Russell and Peter Norvig, 4th edition, Prentice Hall, 2020.
- *Artificial Intelligence: Structures and Strategies for Complex Problem Solving.* George Luger. 5th edition, Addison-Wesley, 2005.
- *Artificial Intelligence: A New Synthesis.* Nils Nilsson. Morgan Kaufmann, 1998.
- *Artificial Intelligence: Theory and Practice.* Dean, Allen, Aloimonos. Benjamin/Cummings Pub. Co., 1998.
- *Artificial Intelligence.* Patrick Winston, 3rd edition, Addison-Wesley, 1992.
- *Artificial Intelligence.* Rich and Knight. 2nd edition, McGraw Hill, 1991.
- *An Introduction to Multi-Agent Systems.* Michael Wooldridge. Wiley, 2nd edition, 2009.
- *Guide to Health Informatics.* Enrico Coiera, CRC Press, 3rd edition, 2015.
- Various research articles



AI is constantly in the news these days!



A screenshot of a Google search results page for the query "artificial intelligence". The search bar at the top shows the query and the Google logo. Below the search bar, there are tabs for "All", "Images", "News", "Videos", "Books", and "More". The "All" tab is selected. The search results show a featured snippet from TechTarget, which defines artificial intelligence as "the simulation of human intelligence processes by machines, especially computer systems." To the right of the snippet are two images: one of a brain on a circuit board and another of a network diagram. Below the snippet, there is a section titled "People also ask" with four questions: "What is artificial intelligence and example?", "What is AI used in today?", "How can I learn AI?", and "Is there a free AI I can use?". To the right of this section is a box titled "Artificial intelligence" with a brief definition and a link to Wikipedia. At the bottom right, there is a section titled "People also search for" with four small images representing different AI-related topics. A red circle highlights the text "About 1,550,000,000 results (0.35 seconds)" in the top right corner of the search results area.

Google

artificial intelligence

Sign in

SafeSearch

Any time ▾ All results ▾ Advanced Search

About 1,550,000,000 results (0.35 seconds)

Artificial intelligence is the simulation of human intelligence processes by machines, especially computer systems. Specific applications of AI include expert systems, natural language processing, speech recognition and machine vision.

TechTarget
https://www.techtarget.com › definition › AI-Artificial-I...

What is Artificial Intelligence (AI)? Everything You Need to Know

People also ask ▾

What is artificial intelligence and example? ▾

What is AI used in today? ▾

How can I learn AI? ▾

Is there a free AI I can use? ▾

Feedback

Artificial intelligence

Field of study ▾

Artificial intelligence, in its broadest sense, is intelligence exhibited by machines, particularly computer systems. [Wikipedia](#)

People also search for



Sooner or later, everyone gets a haircut...



Watch this video: <https://www.youtube.com/watch?v=IXUQ-DdSDoE>

“The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find **how to make machines use language**, **form abstractions and concepts**, **solve kinds of problems now reserved for humans**, and **improve themselves**.”

(From John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon, Proposal for the 1956 Dartmouth Workshop on AI, 1955)



So...what is AI?

Is it about Robots?

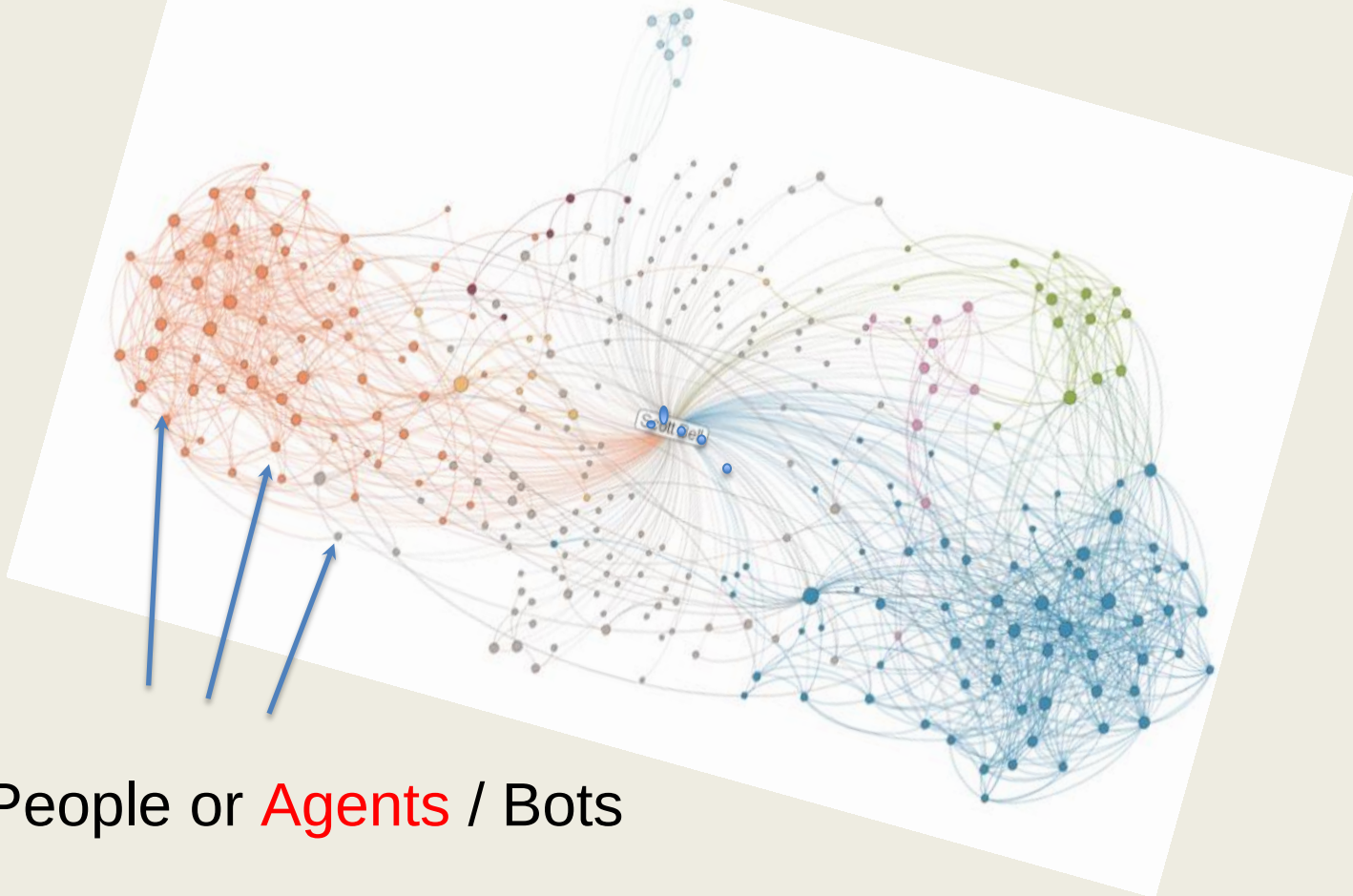
Or Robotic Waiters/Waitresses?

- ...the “strong AI” vision...





It's not a brain... it's a social network... part of the WWW... or is it our collective brain?



People or **Agents** / Bots





It's not a brain...it's a social network...part of the WWW... or is it our collective brain?

A Huge Multi-Agent System

Decision Making

Problem Solving

People or **Agents** / Bots





Who are these agents?

• Rational Agent

- Reactive
- Proactive

• Socially-able

- Acts to achieve the **best possible outcome**, given facts/knowledge + informed opinions/beliefs
- Combines multiple abilities in the best possible manner

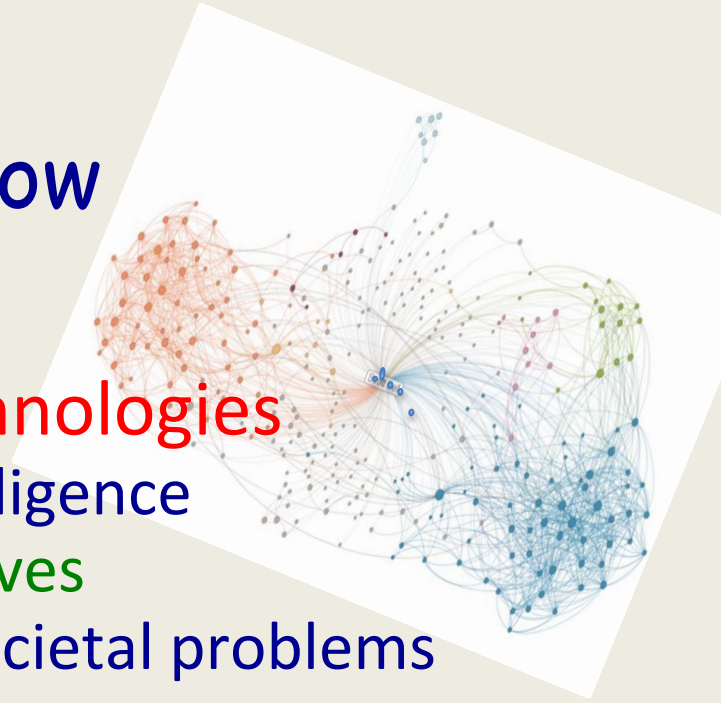
"We are social beings"

Aristotle



AI in the Web of Tomorrow

- **Man-Agents collaboration mediated by digital interaction technologies**
 - combining human **and** machine intelligence
 - **exploiting diverse large-scale collectives**
 - ...to solve complex computational/societal problems
- **Examples:**
 - optimizing the transportation system of a city
 - managing the treatment plans of a patient group
 - coordinating communities of
 - software developers;
 - activists;
 - Smart Grid consumers



Putting the human (back) in the loop?

- In a sense, this is yet another step towards abandoning the “strong AI” vision...



Thus, it's ok to not build the super-human,
omnipotent intelligent agent ...

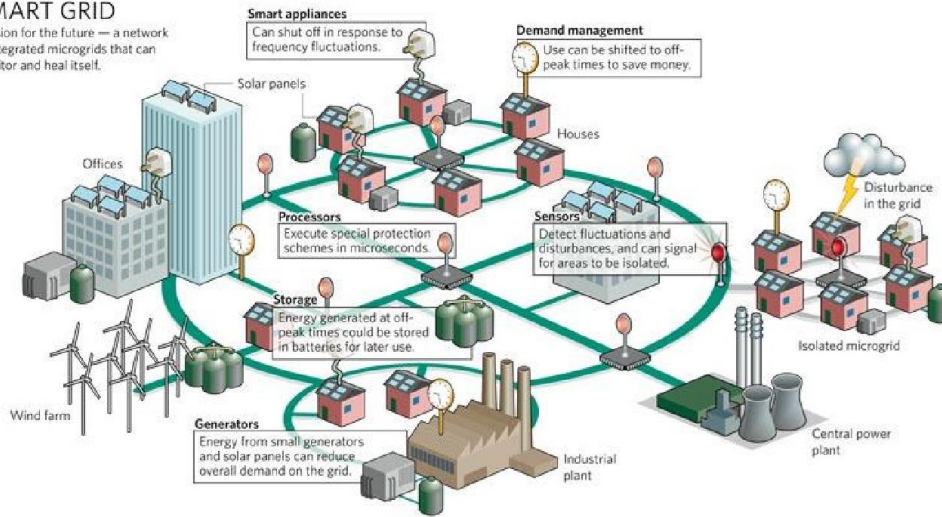


...as long as we get to get things done!...

- ...to ease our everyday living...and even save lives!

SMART GRID

A vision for the future — a network of integrated microgrids that can monitor and heal itself.



AI Applications

- Applications include (*but are not restricted to*) :
 - Medical Support Systems
 - Machine Learning-Based Medical Imaging
 - Decision Support Systems / Decision networks
 - Smart Electricity Grid, Energy efficiency, Virtual Power Plants
 - Sensor networks / wide area surveillance and security systems
 - Natural Language Processing, Understanding
 - Speech Recognition, Generation, Understanding
 - Autonomous Driving
 - Multi-agent coordination
 - Disaster management systems
 - Decentralized resource allocation
 - Self-organizing agent organizations for task execution
 - Personalized recommendation systems and web applications
 - Automated negotiations
 - Computer Games and Recreation
 - Personalized educational environments
 - e-Democracy – enhancing voters' decision making abilities
 - ...



So, can we really have smart machines? How can we build them?

- We had a glimpse at what AI is/might be/can do/cannot do...
- However:

Intelligence Itself is Difficult to Define

thus

AI is Difficult to Define



(Artificial) Intelligence: more than one dimensions

- Vision: The Intelligent Autonomous Agent
 - Multiple desirable characteristics and abilities



(Artificial) Intelligence: more than one dimensions

- AI is not just Machine Learning
- AI is not just Deep Learning
- AI is not just TensorFlow
- AI is not just ChatGPT, ...

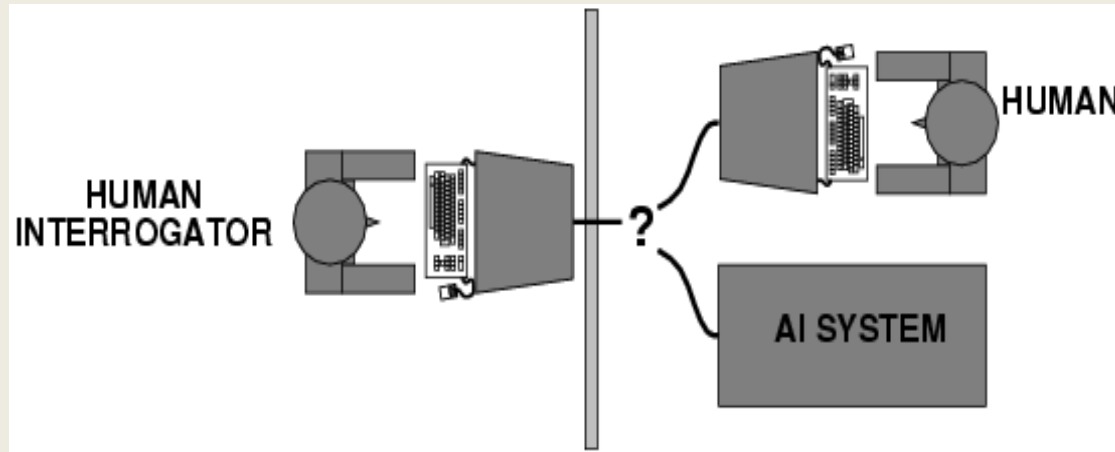


- **Need to combine:** Machine Vision, Machine Learning, Logic, Human-Computer Interaction, Big Data Science, Mathematical Optimization, Decision Theory, Game Theory, Social Choice Theory, Social Anthropology, ...



Can machines be smart?

Turing test (1950)



“Can machines think?”
... Define “machines” ...
... Define “think” ...



Can machines be smart?

- **The “Strong AI” hypothesis:** *if machines act intelligently, as humans do, that means that they also think or that they are smart, like humans are - and we will make such machines*
- Question: *Is “appearances” enough?*
 - (some) Philosophers: *AI simply emulates thinking*
 - **A response:** **expressing similar doubts for a human would have been rude** [Alan Turing]
 - Still, it’s fairly easy and ...a no-brainer to pass a Turing test!
 - *How to Pass a Turing Test...*
 - <http://www.cse.buffalo.edu/~rapaport/Papers/TURING.pdf>
 - “The Chinese Room”



The Chinese Room

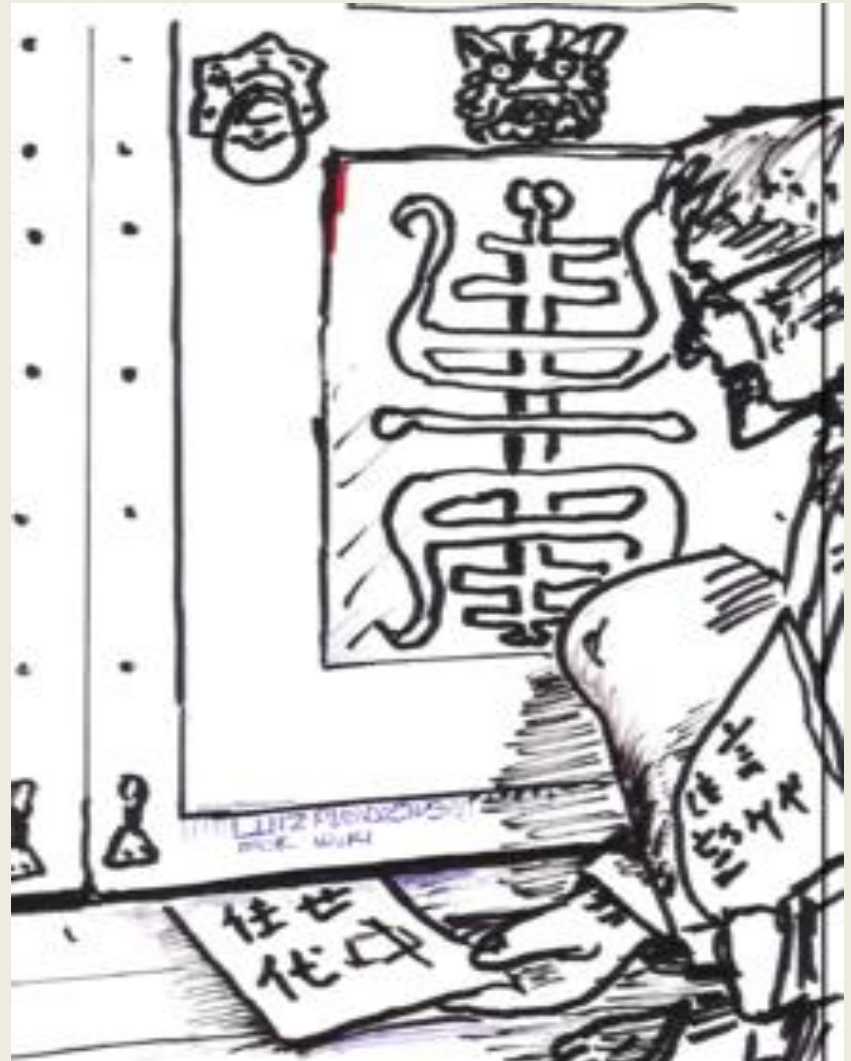


The Chinese Room

- But, ... if a machine uses a series of approximate computations and reaches a response, simply using “math magic” that appears **to have nothing to do with human thinking** (e.g. **probabilistic reasoning**) ... would that make it intelligent/rational?

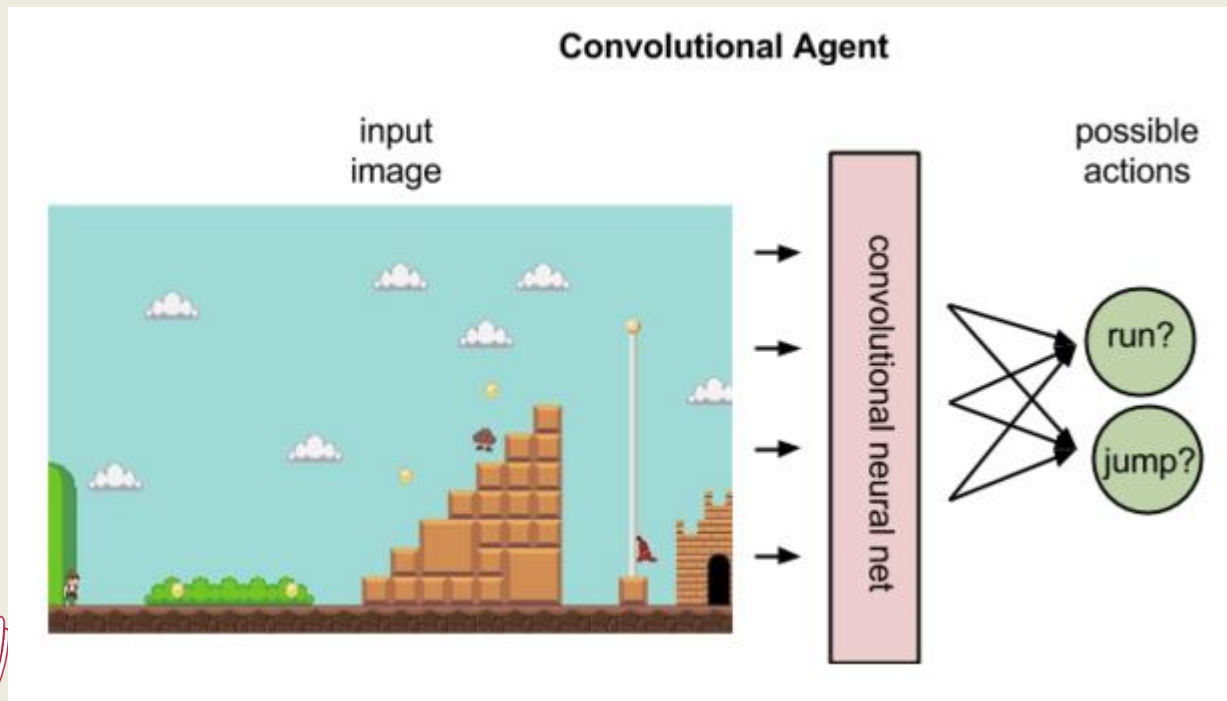
*- **Yes!** There are more possible combinations (or lemmas in the lexicon) than atoms in the universe! If that is not intelligent, then what is?*

*Judea Pearl, Turing Award Speech, 2012
(a liberal rendition by GC)*



Can machines be smart?

- Machines can only *follow (logical) rules*, they cannot be intelligent in the way humans are. [Dreyfus 1972, 1986, 1992]
 - Well, AI has moved far beyond that
 - machines nowadays can learn, even without a teacher



source:
<https://shivam5.github.io/drl/>



Can machines be smart?

- Dreyfus again: The chess grand master argument
 - The right move just “appears” in the mind of the grand master
 - no complex calculations required!
 - Ok...but... *how* does that happen? Philosophers cannot answer the question
 - ...plus grand masters have a tough time with chess-playing machines these days...



Success Stories

Deep Blue (1997) and Deep Fritz (2003)

“25 Years Ago, Chess Changed Forever When Deep Blue Beat Garry Kasparov”

online news, 2021

“Kasparov 'forced' to draw with X3D Fritz”, online news, 2003



Success Stories

Jeopardy!

*“It’s alive: IBM’s Watson
supercomputer
defeats humans in final
Jeopardy match”
online news, 2011*



Success Stories

Mars Rovers and Helicopter

“Seventh Landing Anniversary on Mars for the Twin NASA Rovers” (2011)



“Hilltop Panorama marks Mars Rover’s 11th Anniversary” (2015)

“NASA’s Ingenuity Mars Helicopter to Begin New Demonstration Phase” (2021)



mars.nasa.gov



Success Stories

Companion Robots Care Providing Robots

“Robot nannies play with kids and change their diapers...”

“Japan eyes robotic future for aged care”

“Robot nannies are here, but won’t replace your babysitter – yet”

“Demand for companion robots will hike from \$3.8 billion in 2015 to \$34.1 billion by 2022”

online news, 2010, 2014, 2017



Success Stories

ChatGPT



what do you know about Biomedical Engineering?



Biomedical Engineering (BME) is an interdisciplinary field that applies principles and problem-solving techniques from engineering, biology, and medicine to design and create equipment, devices, computer systems, and software used in healthcare. The goal of BME is to improve patient care by developing technologies that can prevent, diagnose, and treat medical conditions.

Key Areas of Biomedical Engineering:

1. Medical Imaging:



...sadly, not always...

Collateral Losses

“New massacres of civilians in Afghanistan in the last 24 hours with another 12 “collateral losses” to be added to the long list of victims ...”

online news, 2010



Defining AI



The Intelligent Autonomous Agent Paradigm

	human	rationality
thinking	Thinking like humans	Thinking rationally
acting	Acting like humans	Acting rationally

source: the AIMA book – by Russell and Norvig



Human Thinking / Human cognition

- Cognitive (neuro)science
 - Psychology + computer science: study information processing in the brain
 - cognition: trace reasoning steps to brain centers / functions
 - Develop a theory of the human mind
- Confirmation via:
 - psychological / medical experiments on humans
 - interpretation of experiments, predicting subjects' behaviour (top-down)
 - direct identification from neurological data (bottom-up)
- GPS (General Problem Solver): reason like a human
 - Newell and Simon, 1961



Thinking rationally / Logical thinking

- Laws of thought / laws of logic
 - Aristotelian reasoning (“...therefore, Socrates is a mortal”)
 - Theory of Logic, axioms, formal reasoning, proofs
- Logical AI
 - knowledge representation via sentences in some formal logic
 - logical reasoning = thinking
 - any problem can be solved if appropriately represented
- Obstacles
 - cannot adequately (or at all) represent/capture uncertainty or informal knowledge
 - *what’s the difference between theory and practice? ...well...*
 - *in theory there is no difference, in practice there is!*
 - insufficient computational resources to represent/solve problems formally
 - inefficient algorithms for effectively tackling large real-world problems



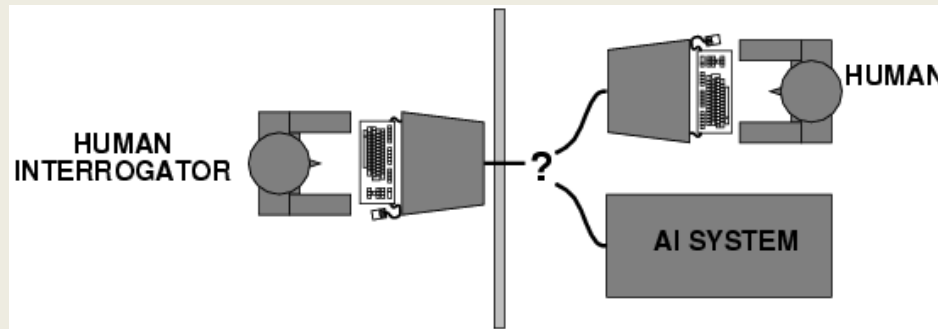
Acting like a human

...do it like a human does...

- Emulating human behavior
 - irrespectively of internal processing
 - focus on successful emulation, not methods/tools/processes used



Turing Test ++



- Natural language processing, knowledge representation, automated reasoning, machine learning ... perhaps computer vision and robotics...
- Yes, but ... does an airplane get mistaken for a bird? Which one flies better? Do their looks or styles matter?



Rational Action / Rational Behaviour

- Autonomous Agent(s)
 - agent = an entity that perceives and acts
 - rational agents
 - Do the Right Thing: perform an action (plan) whose outcome is best in expectation
- Rational Action
 - Define what is “the right thing”/ “best” in some objective manner
 - Rational reasoning: useful, but not a must
 - Some times, only using good reflexes is enough
 - Aim is to succeed in meeting goals and maximizing returns, not on methods/tools/processes used



Which definition is the best?

	human	rationality
thinking	Thinking like humans	Thinking Rationally
acting	Acting like humans	Acting rationally



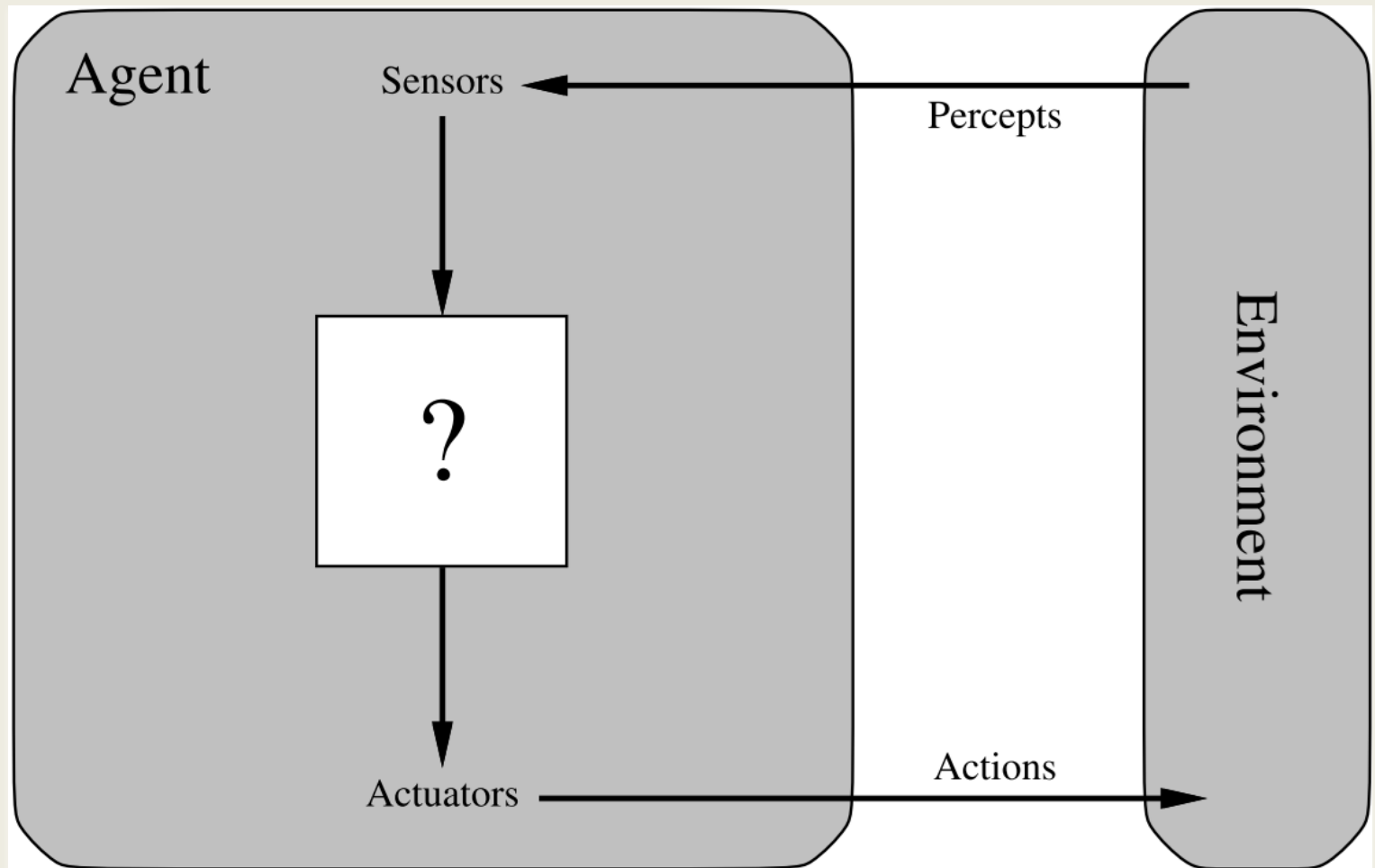
The intelligence we care about

- The “Weak AI” hypothesis:

***We do not care about machines that “think”,
it is enough if they behave “as if they are intelligent”,
that is they choose the “best” actions based on what
they know - in other words, “rationally”***



The Intelligent Autonomous Agent Paradigm ++

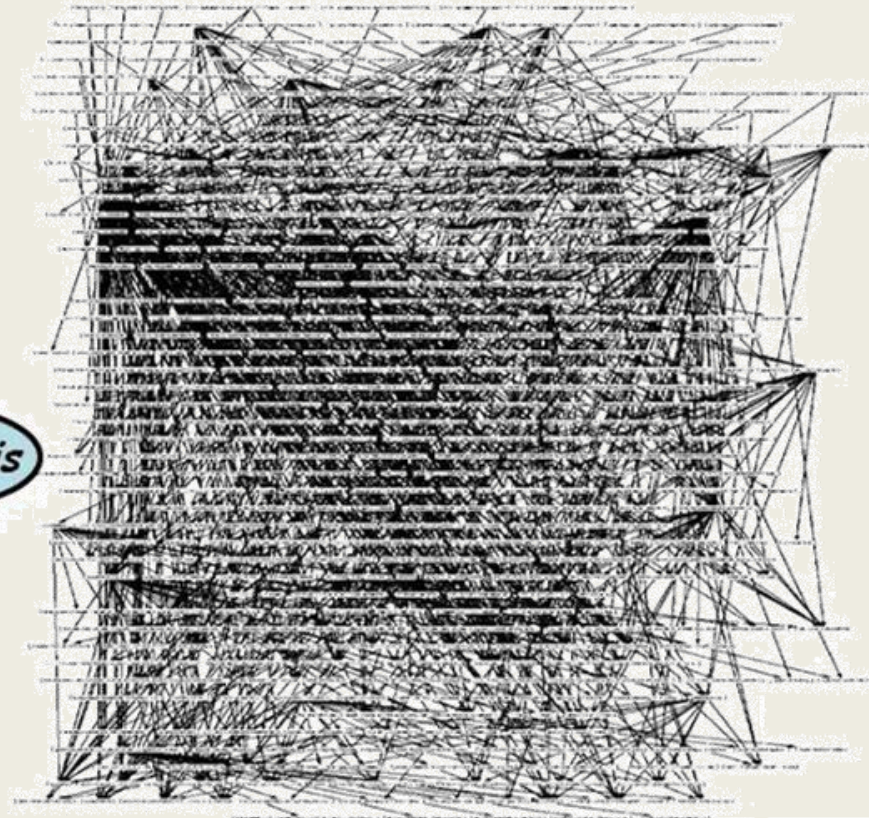
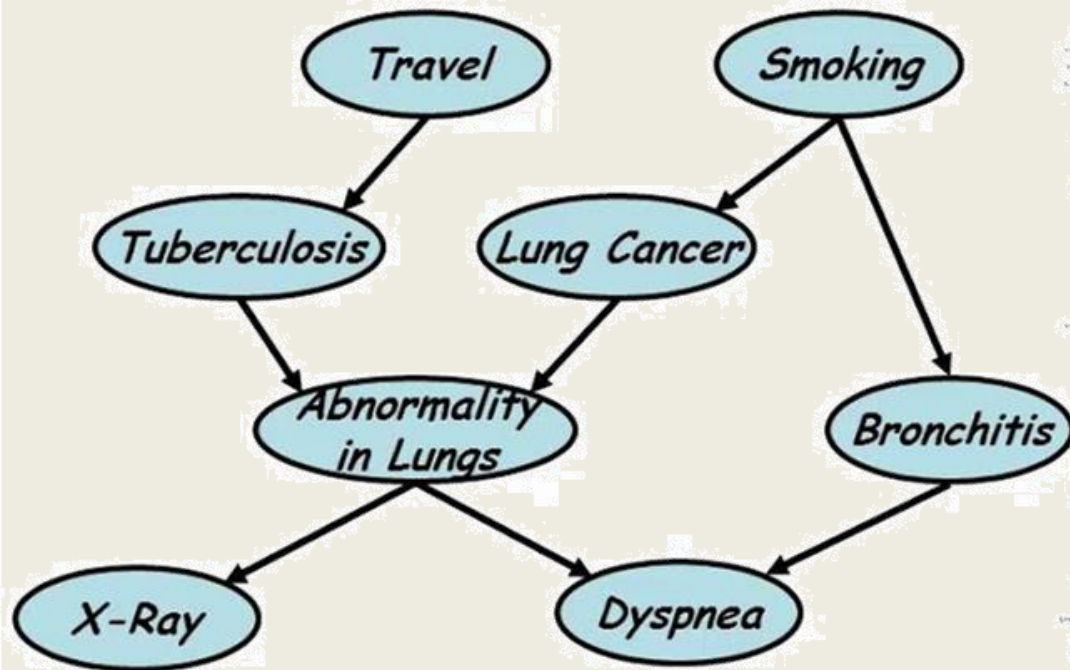


Some More Jargon and Terminology

- Symbolic AI
 - Logic-based + constraints-based AI
 - “deductive thought”:
 - deduces conclusions given “knowledge” and constraints
- Expert Systems
 - emulate experts’ opinions
 - use of appropriately represented built-in knowledge
- Statistical AI
 - probabilistic graphical models (“model-based AI”)
 - crunches numbers and data
 - “inductive thought”: discovers patterns and trends in data
- Data-Driven AI
 - statistical AI + neural networks



Graphical Models



Origins



Contributions to AI

Different fields have contributed to AI in the form of ideas, viewpoints and techniques.

- *Philosophy*: logic, reasoning, mind as a physical system
- *Mathematics*: formal representation and proof algorithms
- *Economics*: formal theory of rational decisions, game theory
- *Neurosciences*: physical substrate for mental activities
- *Psychology*: adaptation, phenomena of perception and motor control
- *Linguistics*: knowledge representation, grammars
- *Computer Engineering*: development of efficient computers
- *Control theory*: homeostatic systems, stability, optimal agent design



Philosophy (428 b.c. – today)

- **Are there any formal rules for drawing conclusions?**
 - Aristotle: exact set of laws for rational intelligence
- **How does intelligence come from the brain?**
 - materialism: intelligence is subject to the laws of nature
 - dualism: part of intelligence is outside the physical world
- **Where does knowledge come from?**
 - empiricism, logical positivism, theory of affirmation
- **How does knowledge lead to action?**
 - logical connection between objectives and knowledge of the effects of actions
 - Aristotle: first planning algorithm with backtracking!



Mathematics (c. 800 – today)

- **What are the typical rules for drawing conclusions?**
 - logic - George Boole (1847)
 - first-order logic - Gottlob Frege (1879)
- **What can be calculated and what not?**
 - Incompleteness Theorem - Godel (1931)
 - Algorithms, Theory of Computation, Solvability - Turing (1936)
 - NP-completeness - Cook (1971), Karp (1972)
- **How do we think with uncertain information?**
 - probability theory - Cardano, Fermat, Pascal, Bernoulli, Laplace
 - the Bayes rule - Thomas Bayes (1702-1761)



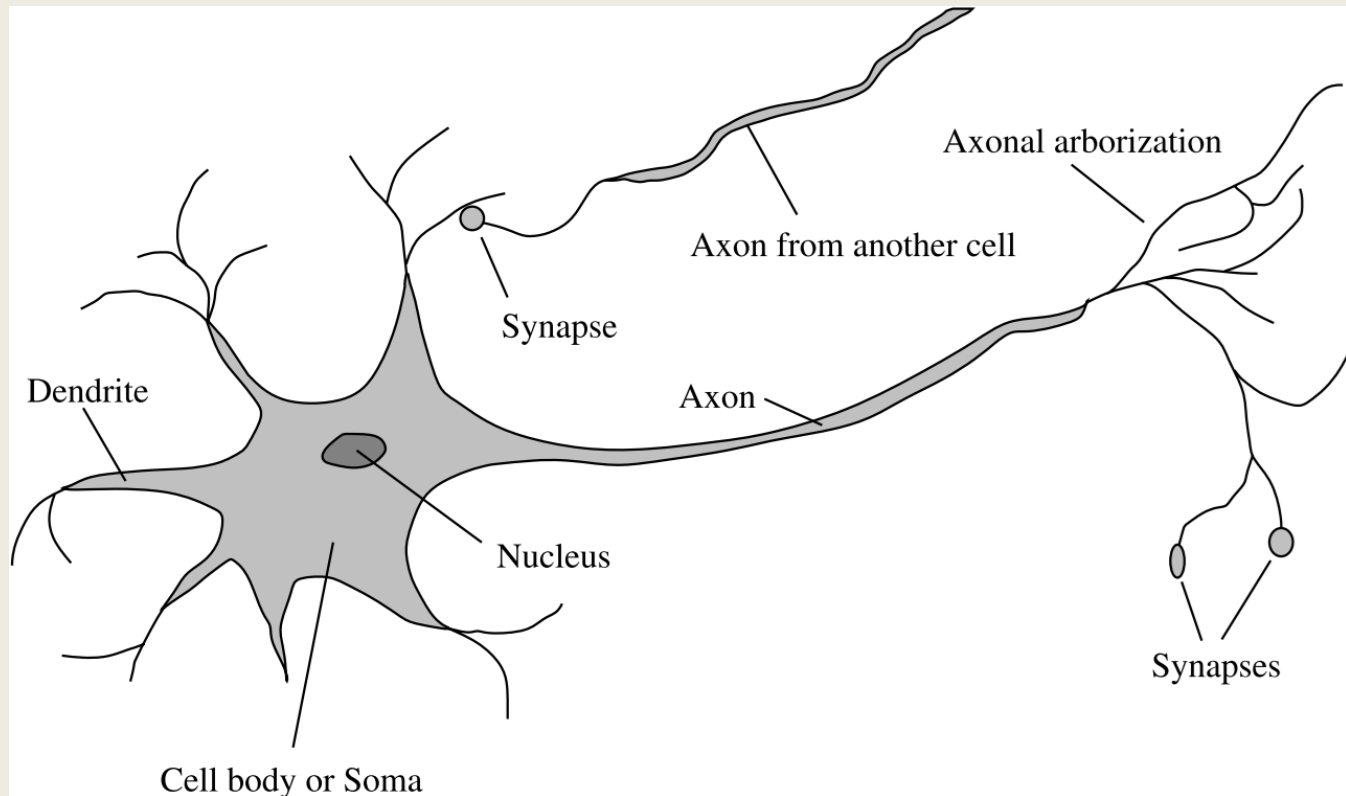
Economics (1776 - today)

- **How do we decide to maximize earnings?**
 - the concept of utility
 - decision theory = probability + utility
- **How do we decide when there is competition?**
 - game theory
- **How do we decide with a view to future earnings?**
 - operations research
 - Markov Decision Processes (MDP) - Bellman (1957)



Neurosciences (1861 - today)

- **How does the brain process information?**
 - a collection of neurons leading to intelligence (?)



Neurosciences (1861 - today)

	Computer 2003	Human Brain
Computational Units	1 CPU, 10^8 gates	10^{11} neurons
Storage Units	10^{10} bit RAM 10^{11} bit disk	10^{11} neurons 10^{14} synapses
Cycle Time	10^{-9} sec	10^{-3} sec
Bandwidth	10^{10} bit/sec	10^{14} bit/sec
Memory Updates/second	10^9	10^{14}

- the computer is 1,000,000 times faster than the brain
- the brain produces 100,000 more work due to parallelism
- there is also a difference in power consumption



Psychology (1879 - today)

- **How do humans and animals think and act?**
 - experimental psychology
 - body endoscopy
 - behaviorism
 - only input - output (stimuli - responses)
 - cognitive psychology
 - stimulus / representation / processing / action sequence
 - information processing device, first forms of agent
 - evolution to cognitive science
 - psychology and computational models
 - "a cognitive theory should look like a program"



Linguistics (1957 - today)

- **How does language relate to thought?**
 - Behavioral Language Learning - Skinner (1957)
 - syntactic language models - Chomsky (1950)
 - computational linguistics
 - natural language processing
 - natural language understanding: complex problem
 - understanding the structure of sentences
 - understanding of the subject
 - understanding of the meaning of the sentences
 - knowledge representation
 - how to represent knowledge to let the computer reason?



Computer Engineering (1940 - today)

- **How do we build an efficient computer?**
 - AI = intelligence + artifact
 - rapid development of efficient computers
 - mutual relationship between CE and AI
 - time sharing
 - interactive interpreters
 - symbolic programming
 - functional programming



Control Theory (1948 - today)

- **How can artifacts be self-controlled?**
 - hydraulic clock with regulator (Ktisivios of Alexandria, 250 BC)
 - thermostat (Drebbel, approx. 1600)
 - modifying behavior when data changes in the environment
 - control theory and government - Weiner (1894-1964)
 - intentional behavior: error minimization mechanism
 - stochastic optimal control
 - objective function maximization
 - difference between AI and modern control theory (CT)
 - AI: discrete systems (natural language, games, vision, ...)
 - CT: continuous systems (mechanical arms, electric current, ...)
 - AI: study / combination of many new fields



A Brief History of AI



Compact History of Artificial Intelligence

1943-1955	The gestation of AI
1956	The birth of AI
1952-1969	Early enthusiasm and high expectations
1966-1973	A dose of reality
1969-1979	Knowledge-based systems
1980-today	AI becomes an industry
1986-today	The return of neural networks
1987-today	AI becomes a science
1995-today	The emergence of intelligent agents
2010-today	Deep learning, ChatGPT, ...



History of Artificial Intelligence

- **1943-1955: The gestation of AI**

- 1943: artificial neural networks (McCulloch and Pitts)
- 1949: learning method for neural synapses (Hebb)
- 1950: “Computing Machinery and Intelligence” (Turing)
 - Turing test, machine learning, genetic algorithms
- 1951: first neural computer (Minsky and Edmonds)



History of Artificial Intelligence ++

- **1956: The birth of AI**

- Two-month workshop at Dartmouth College
 - John McCarthy, Marvin Minsky, Claude Shannon, Nathaniel Rochester, Trenchard More, Arthur Samuel, Ray Solomonoff, Oliver Selfridge, Allen Newell and Herbert Simon
- the Logic Theorist Program (Newell and Simon)
 - proofs for Chapter 2 of Principia Mathematica
- adoption of the term “Artificial Intelligence” (McCarthy)
- AI is studied as a separate scientific field



History of Artificial Intelligence ++

- **1952-1969: Early enthusiasm and high expectations**
 - 1952: a program for Checkers that learns (Samuel)
 - 1958: LISP programming language, Advice Taker (McCarthy)
 - 1959: Geometry Theorem Prover (Gelernter)
 - 1959: AI Lab at MIT (Minsky and McCarthy)
 - 1961: General Problem Solver (Newell and Simon)
 - 1960: Adalines neural networks (Widrow)
 - 1962: Perceptrons (Rosenblatt)
 - 1963: AI Lab at Stanford University (McCarthy)
 - 1963: AI microworlds
 - 1969: Shakey robot at Stanford



History of Artificial Intelligence ++

- **1966-1973: A dose of reality**
 - poor results, “curse of dimensionality”
 - 1966: automatic translation funding in the United States ceases
 - typical example: *"the spirit is willing, but the flesh is weak"*
 - automatic translation: *"the vodka is good, but the meat is rotten"!*
 - 1969: rejection of perceptrons (Minsky and Papert)
 - 1973: AI funding for research in Great Britain ceases
- **1969-1979: Knowledge-based systems**
 - 1969: finding molecular structure - DENDRAL (Buchanan et al.)
 - 1970: Diagnosis of Blood Infections - MYCIN (Stanford)
 - 1970: the Prolog language (logical programming)
 - 1975: frames as a method of representing knowledge



History of Artificial Intelligence ++

- **1980-present: AI becomes an industry**
 - 1982: R1 expert system for DEC sales (McDermott)
 - 40 installed expert systems
 - savings of \$40,000,000 per year (1986)
 - 1981: "Fifth Generation" program (Japan)
 - followed by "responses" from the US (MCC) and the UK
 - AI ?? No !!! It's... IKBS (Intelligent Knowledge-Based Systems)
 - end of 80's: the "winter" of AI due to many promises
- **1986-present: The return of neural networks**
 - 1982: Statistical engineering techniques for the analysis of neural network properties (Hopfield)
 - mid-1980s: Back-propagation (1969), applications in Computer Science and Psychology [Parallel Distributed Processing, 1986]



History of Artificial Intelligence ++

- **1987-present: AI becomes a science**
 - revolution in content and methodology
 - achievements are based on rigorous theories
 - acceptance only after proof or rigorous experimentation
 - cooperation with other fields
 - Hidden Markov Models, Data mining, Bayesian networks, Machine learning, Game theory
- **1995-present: The emergence of intelligent agents**
 - consolidation of all sub-sectors
- **2010-present: Deep Learning**
 - deep neural networks
- **2022-present: ChatGPT**
 - large language models



More AI Successes

- **Medical Diagnosis and Decision Support Systems**

- Expert systems, Bayesian network-based, neural network-based systems

- **Logistics Planning**

- During the Persian Gulf War (1991), the United States used the D.A.R.T. (Dynamic Analysis and Replanning Tool) for automated planning and scheduling of actions for their transfers.

- **Planning**

- Remote Agent: creating plans for NASA Exploration Vehicles (2000)

- **Games**

- AlphaGo defeats world Go champion in Korea (2016)

- **Robotics**

- Stanley autonomously drove 211 km in the desert in 10 hours (2005)

- **Crossword Puzzles**

- Proverb: automatic crossword puzzle solving with standard cues (1999)

- **Protein Folding**

- AlphaFold: a solution to a 50-year-old grand challenge in biology (2020)



Moral Questions

- **AI in the Defense industry: Killer Robots ?**
 - Need for principles to guide / restrain AI weapons research
- **Robots taking over...**
 - *The Matrix...*
 - Stephen Hawking: *“AI could spell the end of human race”*
- **Yet...**
 - “We have to colonize another planet within 100 years”
(again Stephen Hawking)
 - *And thus we need AI / Robots!...*
- **In which case...**
 - *“need for a robots civil rights movement”* (Spielberg’s AI movie)



An Experts' Debate on Intelligence

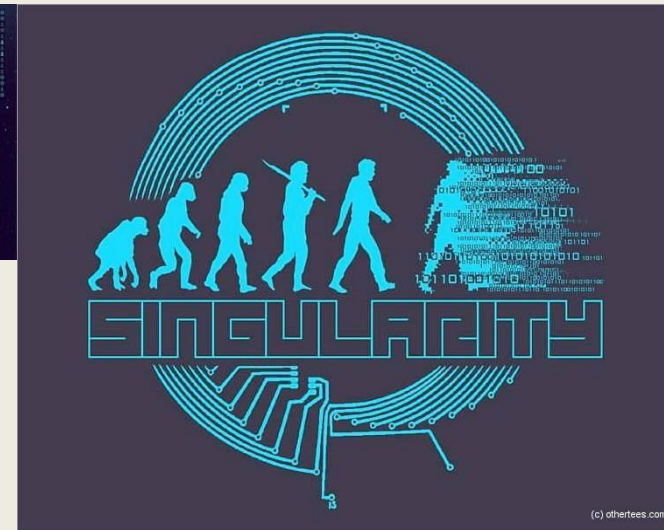
- Noam Chomsky vs. statistical / probabilistic AI
 - Chomsky's critique: *"The field's heavy use of statistical techniques to pick regularities in masses of data is behavioral in nature, and unlikely to yield the explanatory insight that science ought to offer. The "new AI" -- focused on using statistical learning techniques to better mine and predict data - - is unlikely to yield general principles about the nature of intelligent beings or about cognition."*

Brains, Minds and Machines Symposium, MIT, May 2011

- A response by Stuart Russell: *"AI's new methods and definition of progress is not far off from what happens in the other sciences."*



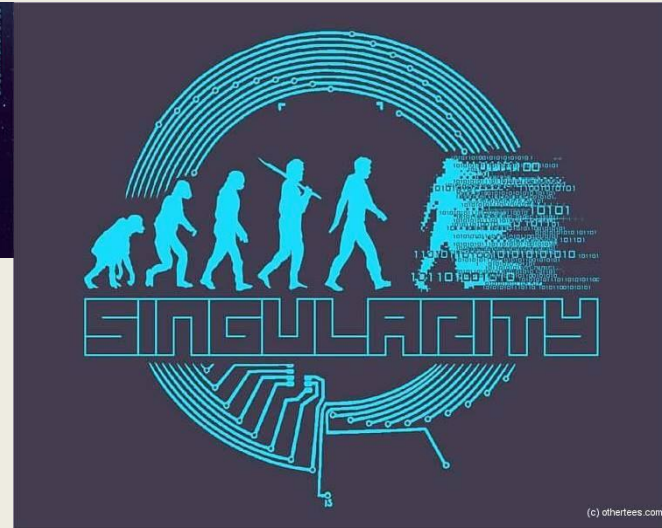
Will a “singularity” emerge?



Maybe we are not “very” far away from building AI
that surpasses human intelligence...
...maybe it’s just the next evolutionary stage...
...maybe not so bad given how human intelligence evolves...



Will a “singularity” emerge?



Maybe we are not “very” far away from building AI that surpasses human intelligence...

...maybe circa 2050 (in ~30 years) *

...not necessarily a bad thing...

...unless we are ready to tackle it ; we should at least *anticipate it*

* maybe not...maybe in 500 years...many questions regarding the computational power necessary and the energy that an AGI (artificial general intelligence) system would require are still open; thus, no good estimate / prediction exists!



The Asilomar AI principles

- 1) **Research Goal:** The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.
- 2) **Research Funding:** Investments in AI should be accompanied by funding for research on ensuring its beneficial use, including thorny questions in computer science, economics, law, ethics, and social studies, such as:
 - How can we make future AI systems highly robust, so that they do what we want without malfunctioning or getting hacked?
 - How can we grow our prosperity through automation while maintaining people's resources and purpose?
 - How can we update our legal systems to be more fair and efficient, to keep pace with AI, and to manage the risks associated with AI?
 - What set of values should AI be aligned with, and what legal and ethical status should it have?



The Asilomar AI principles

- 3) **Science-Policy Link:** There should be constructive and healthy exchange between AI researchers and policy-makers.
- 4) **Research Culture:** A culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI.
- 5) **Race Avoidance:** Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.
- 6) **Safety:** AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.



The Asilomar AI principles

- 7) **Failure Transparency:** If an AI system causes harm, it should be possible to ascertain why.
- 8) **Judicial Transparency:** Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.
- 9) **Responsibility:** Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.
- 10) **Value Alignment:** Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.



The Asilomar AI principles

- 11) **Human Values:** AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.
- 12) **Personal Privacy:** People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.
- 13) **Liberty and Privacy:** The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.
- 14) **Shared Benefit:** AI technologies should benefit and empower as many people as possible.
- 15) **Shared Prosperity:** The economic prosperity created by AI should be shared broadly, to benefit all of humanity.



The Asilomar AI principles

- 16) **Human Control:** Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.
- 17) **Non-subversion:** The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.
- 18) **AI Arms Race:** An arms race in lethal autonomous weapons should be avoided.
- 19) **Capability Caution:** There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.



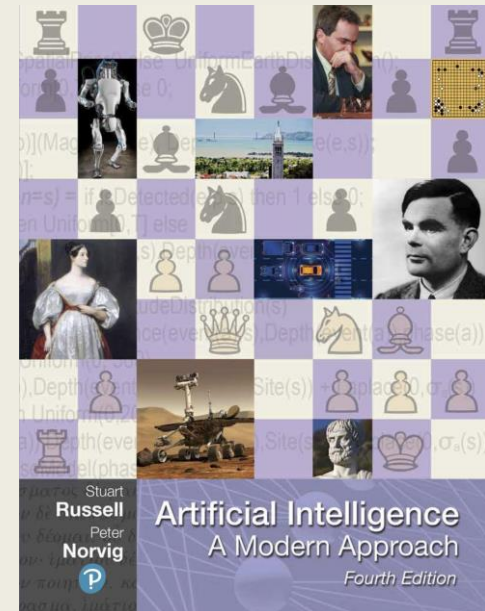
The Asilomar AI principles

- 20) **Importance:** Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.
- 21) **Risks:** Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.
- 22) **Recursive Self-Improvement:** AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.
- 23) **Common Good:** Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.



Suggested readings

- Artificial Intelligence: A Modern Approach,
Stuart Russell and Peter Norvig,
4th edition, Prentice Hall, 2020.
 - Chapter 1



Suggested viewings and listenings

- [Stephen Wolfram on Intelligence, AI, the meaning of life, and more](#)
- [Michael Wooldridge on “The Evolution of AI”](#)
- [Adnan Darwiche on "AI Education" or "AI vs Deep Learning”](#)
- [Michael Wooldridge on “An Introduction to Multiagent Systems”](#)

