

Lecture 08

Artificial Intelligence & Rational Agents

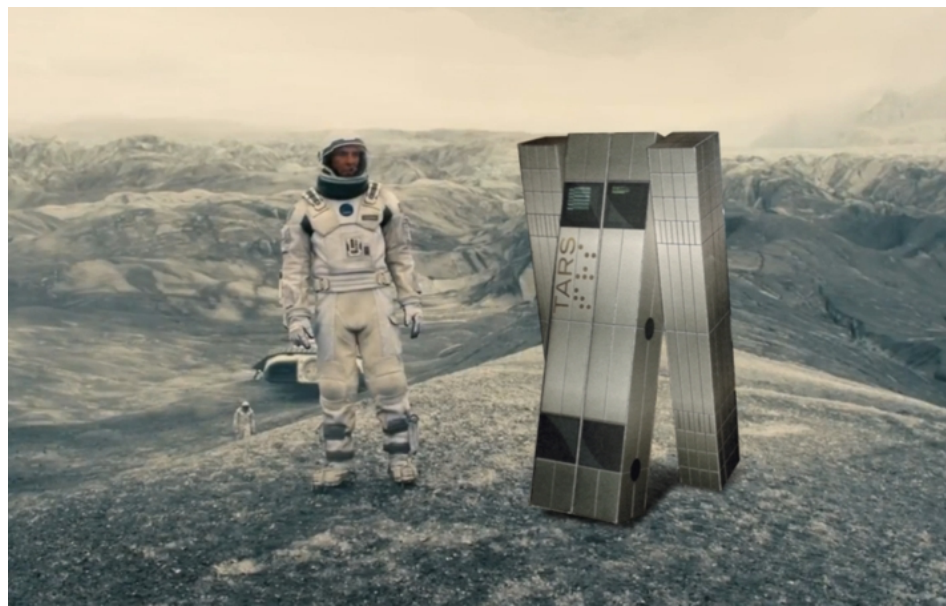
Song Chaoyang

Assistant Professor

Department of Mechanical and Energy Engineering

songcy@sustech.edu.cn

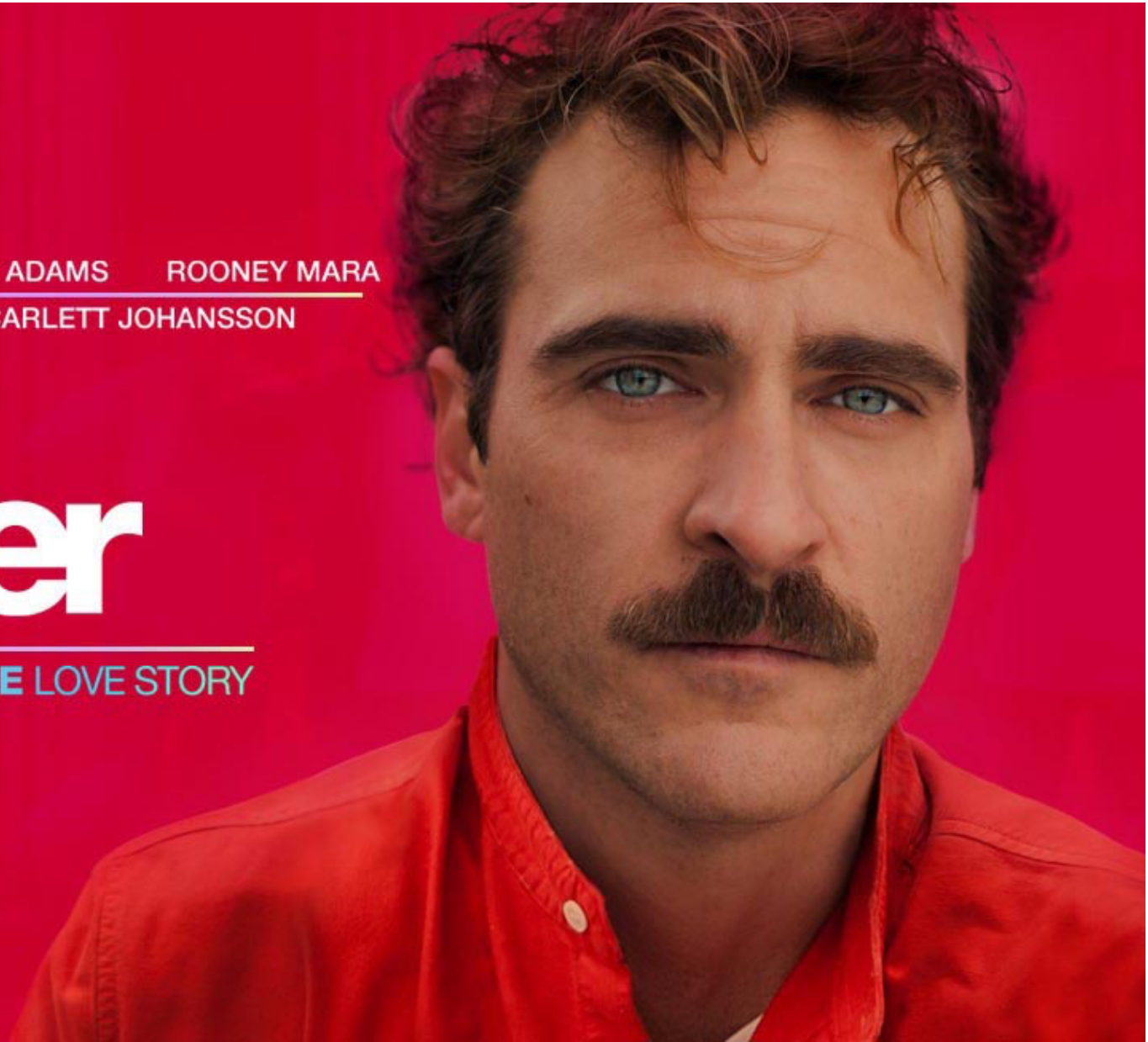




JOAQUIN PHOENIX AMY ADAMS ROONEY MARA
OLIVIA WILDE AND SCARLETT JOHANSSON

her

A SPIKE JONZE LOVE STORY





Artificial Intelligence

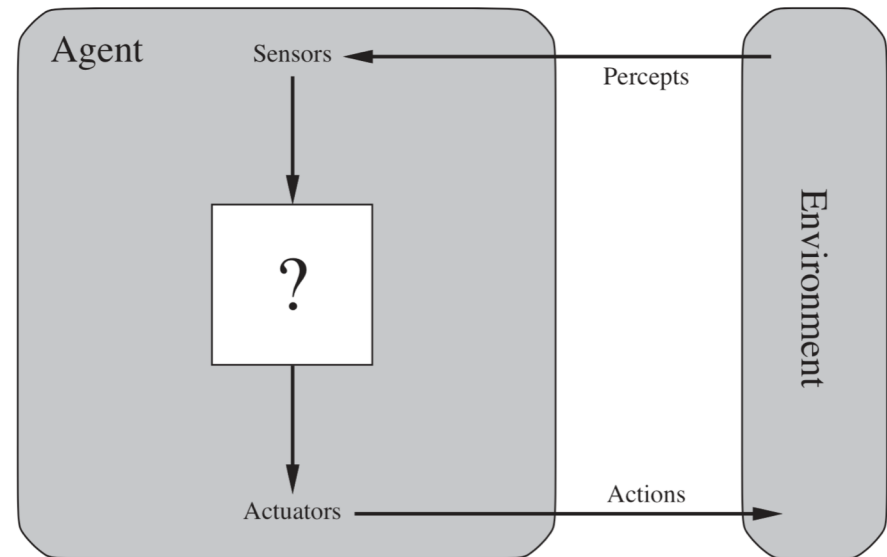
as Computational Rationality

- **Humans** are intelligent to the extent that **our** actions can be expected to achieve **our** objectives
- **Machines** are intelligent to the extent that **their** actions can be expected to achieve **their** objectives
 - Control Theory: minimize cost function
 - Economics: maximize expected utility
 - Operational Research: maximize sum of rewards
 - Statistics: minimize loss function
 - *AI: all of the above, plus logically defined goals*
- **AI \approx Computational Rational Agents**

Designing Rational Agents

Rationality as an *ideal* performance measure

- An **agent** is an entity that *perceives* and *acts*.
 - Intelligence is concerned mainly with **rational action**
- Ideally, an **intelligent agent** takes the best possible action in a situation.
 - A **rational agent** selects actions that maximize its (expected) **utility**.
- Characteristics of the **percepts**, **environment**, and **action space** dictate techniques for selecting rational actions

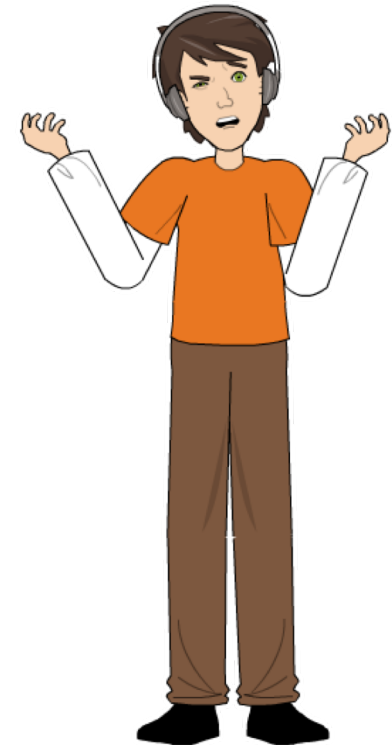


What About the Brain?

“Brains are to intelligence as wings are to flight”

- Brains (human minds) are very good at making rational decisions, but far from perfect; they result from accretion over evolutionary timescales.
- Lessons learned from human minds: memory, knowledge, feature learning, procedure formation, and simulation are key to decision making.

We don't know how they work ...



What is AI?

*Are you concerned with **thinking** or **behavior**?
Do you want to model humans or work from an **ideal standard**?*

Human-centered approach:

observation and hypothesis about human behavior

Thinking Humanly

“The exciting new effort to make computers think . . . *machines with minds*, in the full and literal sense.” (Haugeland, 1985)

“[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning . . .” (Bellman, 1978)

Acting Humanly

“The art of creating machines that perform functions that require intelligence when performed by people.” (Kurzweil, 1990)

“The study of how to make computers do things at which, at the moment, people are better.” (Rich and Knight, 1991)

Rationalist approach:

a combination of mathematics and engineering

Thinking Rationally

“The study of mental faculties through the use of computational models.”
(Charniak and McDermott, 1985)

“The study of the computations that make it possible to perceive, reason, and act.”
(Winston, 1992)

Acting Rationally

“Computational Intelligence is the study of the design of intelligent agents.” (Poole *et al.*, 1998)

“AI . . . is concerned with intelligent behavior in artifacts.” (Nilsson, 1998)

The Foundations of AI

Philosophy

- Can formal rules be used to draw valid conclusions?
- How does the mind arise from a physical brain?
- Where does knowledge come from?
- How does knowledge lead to action?

Mathematics

- What are the formal rules to draw valid conclusions?
- What can be computed?
- How do we reason with uncertain information?

Economics

- How should we make decisions so as to maximize payoff ?
- How should we do this when others may not go along?
- How should we do this when the payoff may be far in the future?

Near miss (1842)

- Babbage design for universal machine
- Lovelace: “a thinking machine” for “all subjects in the universe.”

AncoraSIR.com

Computer engineering

- How can we build an efficient computer?

Psychology

- How do humans and animals think and act?

Control theory and cybernetics

- How can artifacts operate under their own control?

Linguistics

- How does language relate to thought?

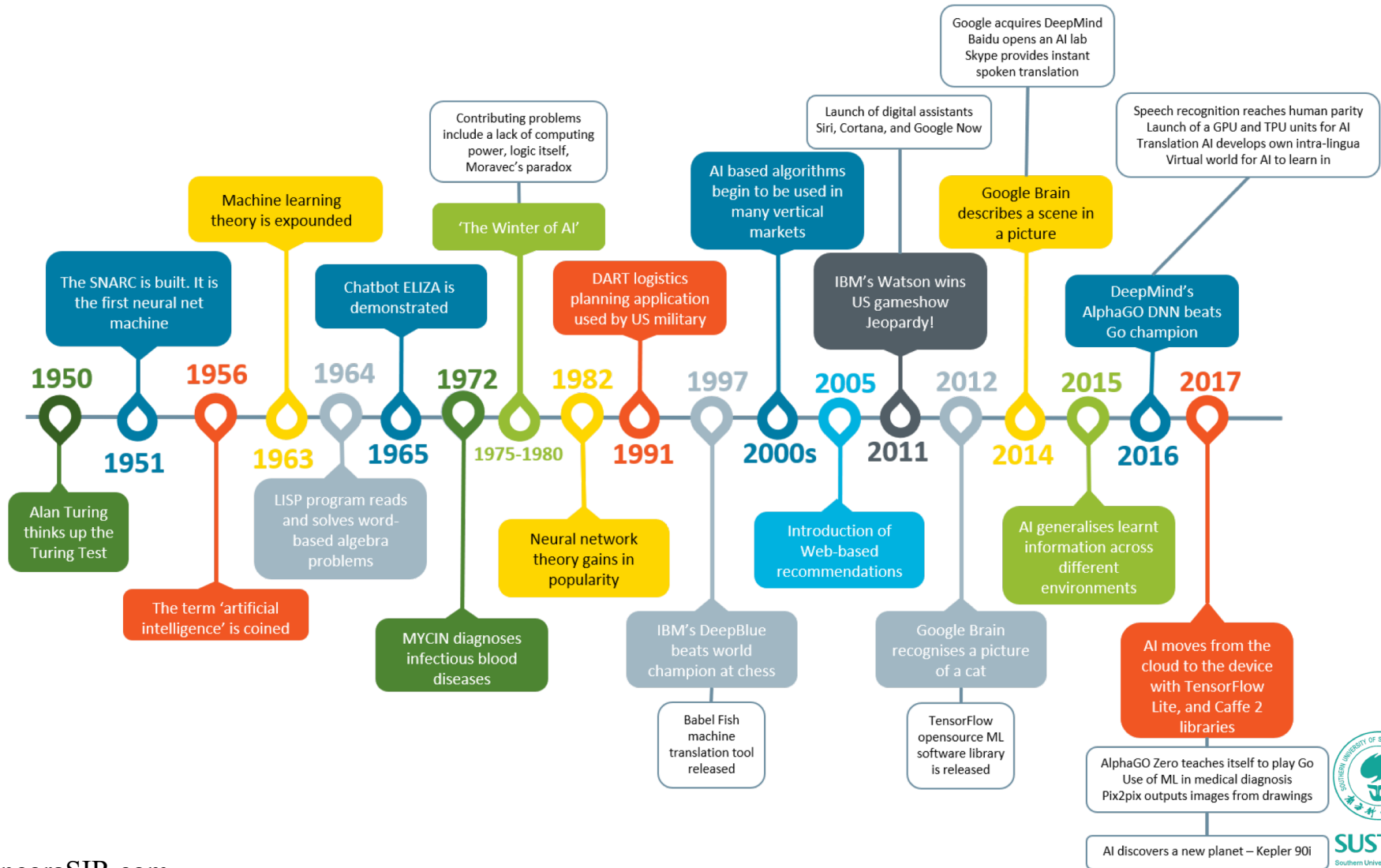
Neuroscience

- How do brains process information?



SUSTech
Southern University
of Science and Technology

A (Short) History of AI

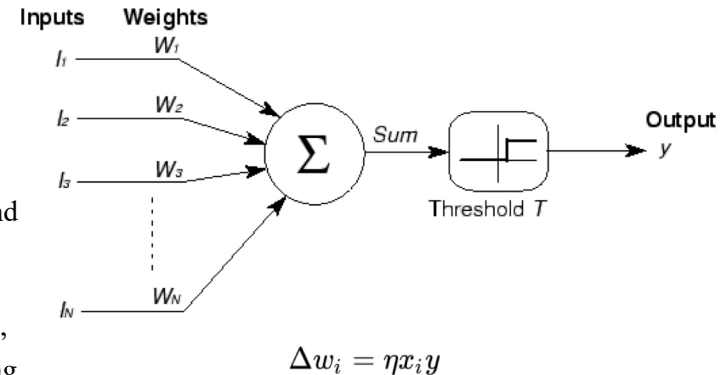


The History of Artificial Intelligence

The gestation of artificial intelligence (1943–1955)

- **Warren McCulloch and Walter Pitts (1943)**

- The first work that is now generally recognized as AI
- Three sources of inspiration
 - knowledge of the basic physiology and function of neurons in the brain;
 - a formal analysis of propositional logic due to Russell and Whitehead; and
 - Turing's theory of computation.
- Proposed a model of artificial neurons
 - each neuron is characterized as being “on” or “off,” with a switch to “on” occurring in response to stimulation by a sufficient number of neighboring neurons.



- **Donald Hebb (1949)**

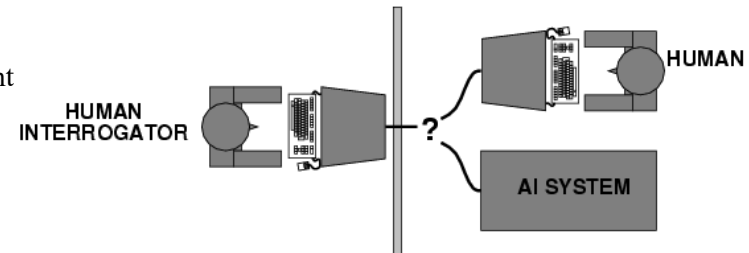
- Hebbian learning, when two joining cells fire simultaneously, the connection between them strengthens.

- **Marvin Minsky and Dean Edmonds (1950)**

- SNARC (Stochastic Neural Analog Reinforcement Calculator), the first neural network computer

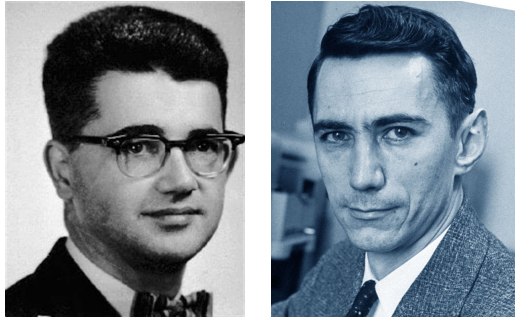
- **Alan Turing (1947~1950)**

- Computing Machinery and Intelligence
 - the Turing Test, machine learning, genetic algorithms, and reinforcement learning
- Child Programme
 - “Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulated the child’s?”



The History of Artificial Intelligence

The birth of artificial intelligence (1956)



John McCarthy and Claude Shannon Dartmouth Workshop Proposal

“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. *We think that a significant advance can be made if we work on it together for a summer.*”

Why it was necessary for AI to become a separate field

- *AI as a branch of control theory? operations research? decision theory? mathematics?*
- **Philosophy:** AI from the start embraced the idea of duplicating human faculties such as creativity, self-improvement, and language use
- **Methodology:** AI is the only one of these fields that is clearly a branch of computer science (although operations research does share an emphasis on computer simulations), and AI is the only field to attempt to build machines that will function autonomously in complex, changing environments.

The History of Artificial Intelligence

Early enthusiasm, great expectations (1952–1969)

• **Early AI Programs**

- Herbert Gelernter (1959): Geometry Theorem Prover
- Arthur Samuel (1952 onwards): checkers, chess, etc.
- John McCarthy (1958): LISP, Time Sharing, Advice Talker

• **General-purpose Methods for Logical Reasoning**

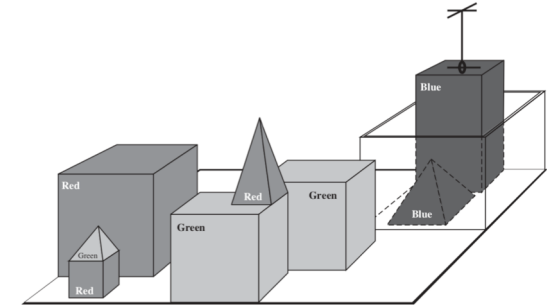
- McCarthy (1965): the ultimate Advice Talker
- J. A. Robinson (1965): a complete theorem-proving algorithm for first-order logic

• **Microworlds Solving Limited Problems**

- James Slagle's SAINT program (1963): solve typical closed-form calculus integration problems
- Tom Evans's ANALOGY program (1968): solved geometric analogy problems that appear in IQ tests.
- Daniel Bobrow's STUDENT program (1967): solved algebra story problems
- The blocks world problem, i.e. to rearrange a set of solid blocks placed on a tabletop in a certain way
 - David Huffman (1971), David Waltz (1975), Patrick Winston (1970), Terry Winograd (1972)

• **Neural Networks based on McCulloch and Pitts (1943)**

- Winograd and Cowan (1963): representation of an individual concept using a large number of elements
- Widrow and Hoff (1960) & Widrow (1962): adalines as an enhancement to Hebb's learning methods
- Frank Rosenblatt (1962): perceptrons
- Block *et al.* (1962): perceptron convergence theorem, the learning algorithm can adjust the connection strengths of a perceptron to match any input data, provided such a match exists



The History of Artificial Intelligence

A dose of reality (1966–1973)

- **Herbert Simon's Prediction in 1957**

- Within 10 years a computer would be
 - chess champion, and
 - a significant mathematical theorem would be proved by machine.
- Came true (or approximately true) within 40 years ☺

It is not my aim to surprise or shock you—but the simplest way I can summarize is to say that there are now in the world machines that think, that learn and that create. Moreover, their ability to do these things is going to increase rapidly until—in a visible future—the range of problems they can handle will be coextensive with the range to which the human mind has been applied.

- **Difficulty #1:** most early programs knew nothing of their subject matter; they succeeded by means of simple syntactic manipulations

- I.e. accurate translation requires background knowledge to resolve ambiguity and establish the content
- “the spirit is willing but the flesh is weak” as “the vodka is good but the meat is rotten”
- “*there has been no machine translation of general scientific text, and none is in immediate prospect.*”

- **Difficulty #2:** the intractability of many of the problems that AI was attempting to solve

- The theory of computational complexity, i.e. problems with the Blocks World
- “scaling up” to larger problems is NOT simply a matter of faster hardware and larger memories
- *The fact that a program can find a solution in principle does not mean that the program contains any of the mechanisms needed to find it in practice.*

- **Difficulty #3:** some fundamental limitations on the basic structures being used to generate intelligent behavior

- Minsky and Papert's book *Perceptrons* (1969): although perceptrons (a simple form of neural network) could be shown to learn anything they were capable of representing, they could represent very little
- Bryson and Ho (1969): proposed back-propagation learning, later used in multilayer NN in 1980s

The History of Artificial Intelligence

Knowledge-based systems: The key to power? (1969–1979)

- The Alternative to **Weak methods**
 - Use more powerful, domain-specific knowledge that allows larger reasoning steps and can more easily handle typically occurring cases in narrow areas of expertise.
 - *“To solve a hard problem, you have to almost know the answer already”*
- **The *Knowledge-Intensive System***
 - Deriving expertise from large numbers of special-purpose rules
 - Heuristic Programming Project (HPP) as an **expert systems**
 - The DENDRAL program for inferring molecular structure
 - Domain specific general theoretical model, and no problem of uncertainty
 - MYCIN as medical diagnosis for blood infection
 - With about 450 rules, MYCIN was able to perform as well as some experts, and considerably better than junior doctors.
 - Roger Schank’s work on natural language processing
 - Representing stereotypical situations (Cullingford, 1981)
 - Describing human memory organization (Rieger, 1976; Kolodner, 1983), and
 - Understanding plans and goals (Wilensky, 1983)

The History of Artificial Intelligence

AI becomes an industry (1980–present) & The return of neural networks (1986–present)

- **To Build Intelligent Programs & Computers**

- Digital Equipment Corporation (DEC): - \$40 mil/yr
 - The first successful commercial expert system, R1, helped configure orders for new computer systems
- “Fifth Generation” project and Microelectronics and Computer Technology Corporation (MCC)
 - Never met the goals ☹
- Million \$ in 1908 => Billion \$ in 1988
- **AI Winter**: Fall by the wayside as they failed to deliver on extravagant promises

- **Connectionist or Symbolic Approach?**

- Whether symbol manipulation had any real explanatory role in detailed models of cognition?
 - General view: complementary, not competing
- Bifurcated paths
 - Creating effective network architectures and algorithms and understanding their mathematical properties
 - Careful modeling of the empirical properties of actual neurons and ensembles of neurons

The History of Artificial Intelligence

AI adopts the scientific method (1987–present)

- **Scientific Approach**

- To be accepted, hypotheses must be subjected to rigorous empirical experiments, and the results must be analyzed statistically for their importance (Cohen, 1995)
 - *More common to build on existing theories than to propose brand-new ones, to base claims on rigorous theorems or hard experimental evidence rather than on intuition, and to show relevance to real-world applications rather than toy examples*
- Now possible to replicate experiments by using shared repositories of test data and code ☺

- **Hidden Markov Models (HMMs) for Speech Recognition**

- As a mathematical framework for understanding the problem and support the engineering claim that they work well in practice
 - based on a rigorous mathematical theory
 - generated by a process of training on a large corpus of real speech data

- **Information Theory for Machine Translation**

- **Bayesian Network for Uncertain Reasoning and Expert Systems**

- **Machine Learning for Robot Learning with Vision**

The History of Artificial Intelligence

The emergence of intelligent agents (1995–present)

- Solving the “**Whole Agent**” Problem, **Again**
 - SOAR (Newell, 1990; Laird *et al.*, 1987)
 - The best-known example of a complete agent architecture.
 - The Internet as an important environments for intelligent agents.
- The Realization of **Reorganizing Subfields** of AI
 - Handle uncertainty, as sensory systems (vision, sonar, speech recognition, etc.) cannot deliver perfectly reliable information about the environment
 - AI has been drawn into much closer contact with other fields
- **Back to its Roots**, Not an Improved Application?
 - **Human-Level AI** as a result of discontent with the progress of AI
 - “Machines that think, that learn and that create”
 - John McCarthy (2007), Marvin Minsky (2007), Nils Nilsson (1995, 2005) and Patrick Winston (Beal and Winston, 2009)
- **Artificial General Intelligence**
 - (Goertzel and Pennachin, 2007)
 - A universal algorithm for learning and acting in any environment
- **Friendly AI**
 - (Yudkowsky, 2008; Omohundro, 2008)

The History of Artificial Intelligence

The availability of very large data sets (2001–present)

- **The Increasing Availability Of Very Large Data Sources**

- Trillions of words of English and billions of images from the Web (Kilgarriff and Grefenstette, 2006);
- Billions of base pairs of genomic sequences (Collins *et al.*, 2003)

- **Word-sense Disambiguation as an Example**

- Yarowsky's (1995)
 - Given the use of the word “plant” in a sentence, does that refer to flora or factory? (Bootstrap)
- Banko and Brill (2001)
 - A mediocre algorithm with 100 million words of unlabeled training data outperforms the best known algorithm with 1 million words

- **The “Knowledge Bottleneck” In AI**

- The problem of how to express all the knowledge that a system needs
 - May be solved in many applications by learning methods rather than hand-coded knowledge engineering, provided the learning algorithms have enough data to go on (Halevy *et al.*, 2009)

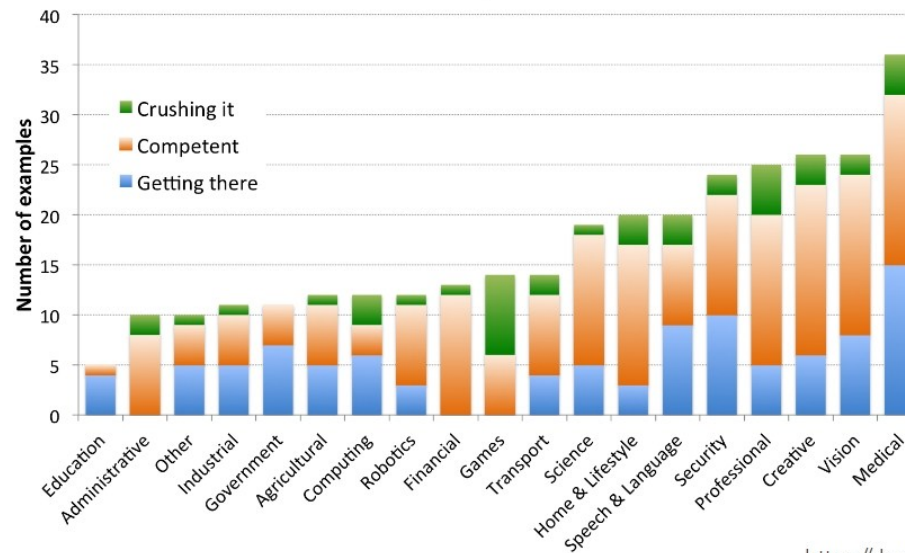
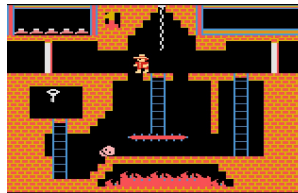
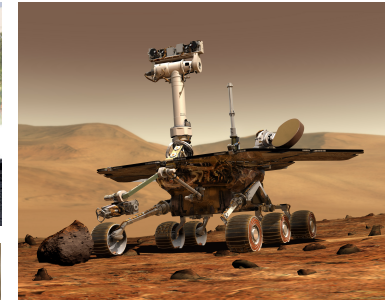


Brynjolfsson, E. & McAfee, A. (2014). The second machine-age: Work, progress, and prosperity in a time of brilliant technologies.

The State of the Art

What can AI do today?

- Robotic vehicles
- Speech recognition
- Autonomous planning and scheduling
- Game playing
- Spam fighting
- Logistics planning
- Robotics
- Machine Translation
- ... Lots More



AncoraSIR.com

<https://deepindex.org>



A Summary of AI

- **Philosophers** (going back to 400 B.C.)
 - made AI conceivable by considering the ideas that the mind is in some ways like a machine, that it operates on knowledge encoded in some internal language, and that thought can be used to choose what actions to take.
- **Mathematicians**
 - provided the tools to manipulate statements of logical certainty as well as uncertain, probabilistic statements. They also set the groundwork for understanding computation and reasoning about algorithms.
- **Economists**
 - formalized the problem of making decisions that maximize the expected outcome to the decision maker.
- **Neuroscientists**
 - discovered some facts about how the brain works and the ways in which it is similar to and different from computers.
- **Psychologists**
 - adopted the idea that humans and animals can be considered information- processing machines. Linguists showed that language use fits into this model.
- **Computer engineers**
 - provided the ever-more-powerful machines that make AI applications possible.
- **Control theory**
 - deals with designing devices that act optimally on the basis of feedback from the environment. Initially, the mathematical tools of control theory were quite different from AI, but the fields are coming closer together.

Cycle of success, misplaced optimism and cutbacks

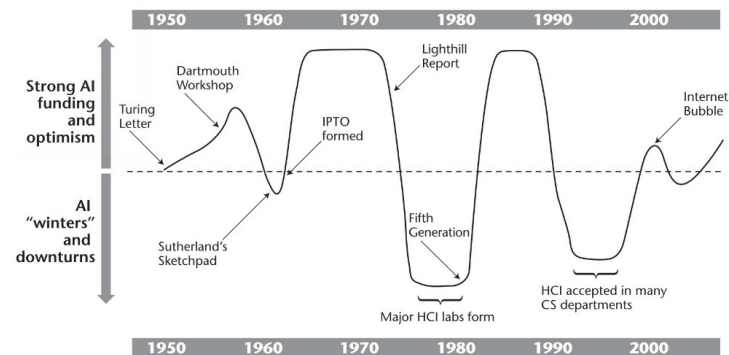
Cycles of introducing new creative approaches and systematically refining the best ones.

Advanced more rapidly in the **past decade**

- greater use of the scientific method in experimenting with and comparing approaches.

Recent progress in understanding the theoretical basis for intelligence has gone hand in hand with improvements in the capabilities of **real systems**.

- The subfields of AI have become more integrated, and AI has found common ground with other disciplines.



Grudin, J. (2009)
AI and HCI: Two fields divided by a common focus.

Frankish, K. & Ramsey, W.M. (2017)
The Cambridge Handbook of Artificial Intelligence.



SUSTech
Southern University
of Science and Technology

Future of AI

- We are doing AI...
 - To create intelligent systems
 - The more intelligent, the better
 - To gain a better understanding of human intelligence
 - To magnify those benefits that flow from it
 - E.g., net present value of human-level AI \geq \$13,500T
 - Might help us avoid war and ecological catastrophes, achieve immortality and expand throughout the universe
- What if we succeed?
 - Still missing**
 - Real understanding of language
 - Integration of learning with knowledge
 - Long-range thinking at multiple levels of abstraction
 - Cumulative discovery of concepts and theories

Date unpredictable

What's Bad about Better AI?

Value Misalignment & Instrumental Goals

- AI that is incredibly good at achieving something other than what we really want
- AI, economics, statistics, operations research, control theory all assume utility to be *exogenously specified*
- For *any primary goal*, the odds of success are improved by
 - 1) Maintaining one's own existence
 - 2) Acquiring more resources
- With value misalignment, these lead to obvious problems for humanity

“We had better be quite sure that the purpose put into the machine is the purpose which we really desire.”

--Norbert Wiener, 1960

Artificial Intelligence

as Computational Rationality

- **Humans** are intelligent to the extent that **our** actions can be expected to achieve **our** objectives
- ~~**Machines** are intelligent to the extent that **their** actions can be expected to achieve **their** objectives~~
 - Control Theory: minimize cost function
 - Economics: maximize expected utility
 - Operational Research: maximize sum of rewards
 - Statistics: minimize loss function
 - *AI: all of the above, plus logically defined goals*
- **Machines** are **beneficial** to the extent that **their** actions can be expected to achieve **our** objectives
 - We need machines to be **provably beneficial**

We don't want machines that are intelligent in this sense

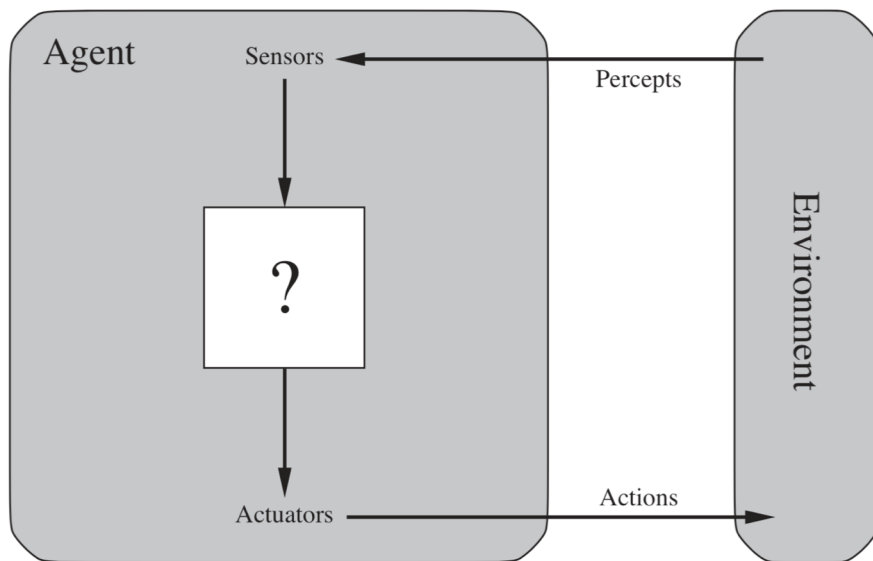
Provably Beneficial AI

*The standard view of AI is a special case,
where the human can exactly and correctly program the objective into the machine*

1. The machine's only objective is to maximize the realization of human preferences
 2. The robot is initially uncertain about what those preferences are
 3. Human behavior provides evidence about human preferences
- **Can we affect the future of AI?**
 - Can we reap the benefits of super-intelligent machines and avoid the risks?
 - ***“The essential task of our age.”***
 - Nick Bostrom, Professor of Philosophy, Oxford University.

Intelligent Agents

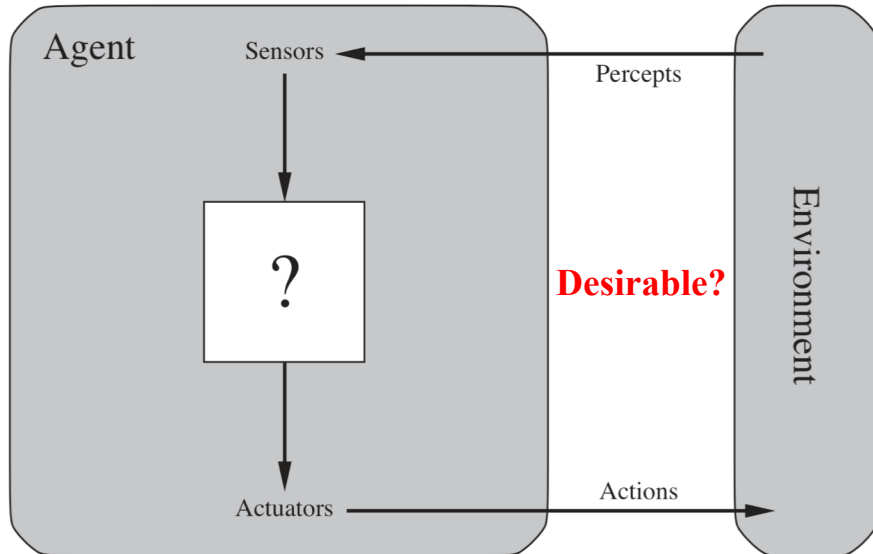
- An **agent** is anything that can be viewed as perceiving its **environment** through **sensors** and acting upon that environment through **actuators**.



The Concept of Rationality

Good Behavior

- A *rational agent* chooses actions to maximize the *expected* utility
 - Today: agents that have a goal, and a cost
 - E.g., reach goal with lowest cost
 - Later: agents that have numerical utilities, rewards, etc.
 - E.g., take actions that maximize total reward over time (e.g., largest profit in \$)



This notion of *desirability* is captured by a **performance measure** that evaluates any given sequence of *environment states*.

The designer's task

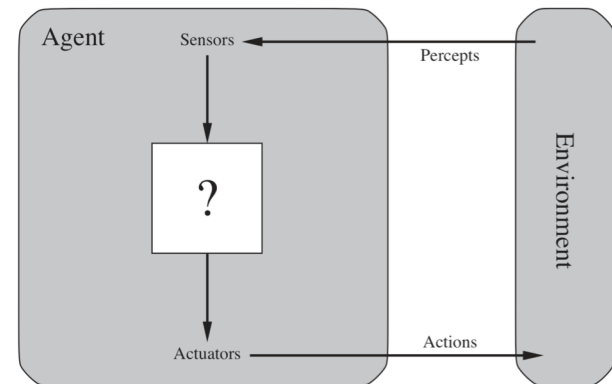
not *agent* states

As a general rule, it is better to design performance measures according to what one actually wants in the environment, rather than according to how one thinks the agent should behave.

Rational Agent

Rationality is not the same as perfection

- Design Factors of a Rational Agent
 - The performance measure that defines the criterion of success.
 - The agent's prior knowledge of the environment.
 - The actions that the agent can perform.
 - The agent's percept sequence to date.
- *For each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has.*
- **Rationality** maximizes *expected* performance,
 - while **perfection** maximizes *actual* performance.



The Nature of Environments

The environment type largely determines the agent design

- **Fully observable vs. partially observable**
 - Agent requires *memory* (internal state)
- **Single agent vs. multiagent**
 - Agent may need to behave *randomly*
- **Deterministic vs. stochastic**
 - Agent may have to prepare for *contingencies*
- **Episodic vs. sequential**
 - Agent may need to consider *consequences*
- **Static vs. dynamic**
 - Agent may need to adapt to *change*
- **Discrete vs. continuous**
 - Agent may not be able to enumerate *all states*
- **Known vs. unknown**
 - Agent may not have the *full knowledge*

Agent Type	Performance Measure	Environment	Actuators	Sensors
Medical diagnosis system	Healthy patient, reduced costs	Patient, hospital, staff	Display of questions, tests, diagnoses, treatments, referrals	Keyboard entry of symptoms, findings, patient's answers
Satellite image analysis system	Correct image categorization	Downlink from orbiting satellite	Display of scene categorization	Color pixel arrays
Part-picking robot	Percentage of parts in correct bins	Conveyor belt with parts; bins	Jointed arm and hand	Camera, joint angle sensors
Refinery controller	Purity, yield, safety	Refinery, operators	Valves, pumps, heaters, displays	Temperature, pressure, chemical sensors
Interactive English tutor	Student's score on test	Set of students, testing agency	Display of exercises, suggestions, corrections	Keyboard entry

Task Environment	Observable	Agents	Deterministic	Episodic	Static	Discrete
Crossword puzzle	Fully	Single	Deterministic	Sequential	Static	Discrete
Chess with a clock	Fully	Multi	Deterministic	Sequential	Semi	Discrete
Poker	Partially	Multi	Stochastic	Sequential	Static	Discrete
Backgammon	Fully	Multi	Stochastic	Sequential	Static	Discrete
Taxi driving	Partially	Multi	Stochastic	Sequential	Dynamic	Continuous
Medical diagnosis	Partially	Single	Stochastic	Sequential	Dynamic	Continuous
Image analysis	Fully	Single	Deterministic	Episodic	Semi	Continuous
Part-picking robot	Partially	Single	Stochastic	Episodic	Dynamic	Continuous
Refinery controller	Partially	Single	Stochastic	Sequential	Dynamic	Continuous
Interactive English tutor	Partially	Multi	Stochastic	Sequential	Dynamic	Discrete

The Structure of Agents

$$\textit{Agent} = \textit{Architecture} + \textit{Program}$$

- **Program** reflects the agent function that maps from percepts to actions.
- **Architecture** allows programs to run on some sort of computing device with physical sensors and actuators

function TABLE-DRIVEN-AGENT(*percept*) **returns** an action

persistent: *percepts*, a sequence, initially empty

table, a table of actions, indexed by percept sequences, initially fully specified

append *percept* to the end of *percepts*

action ← LOOKUP(*percepts*, *table*)

return *action*

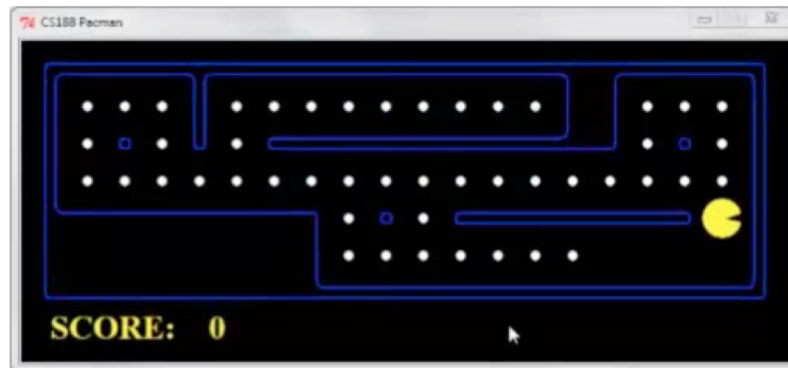
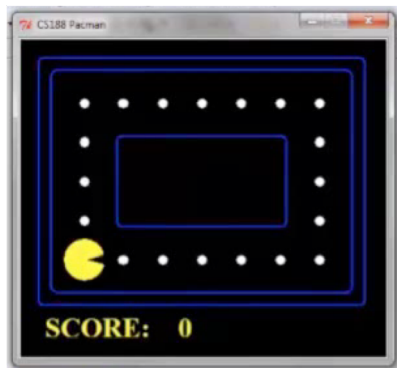
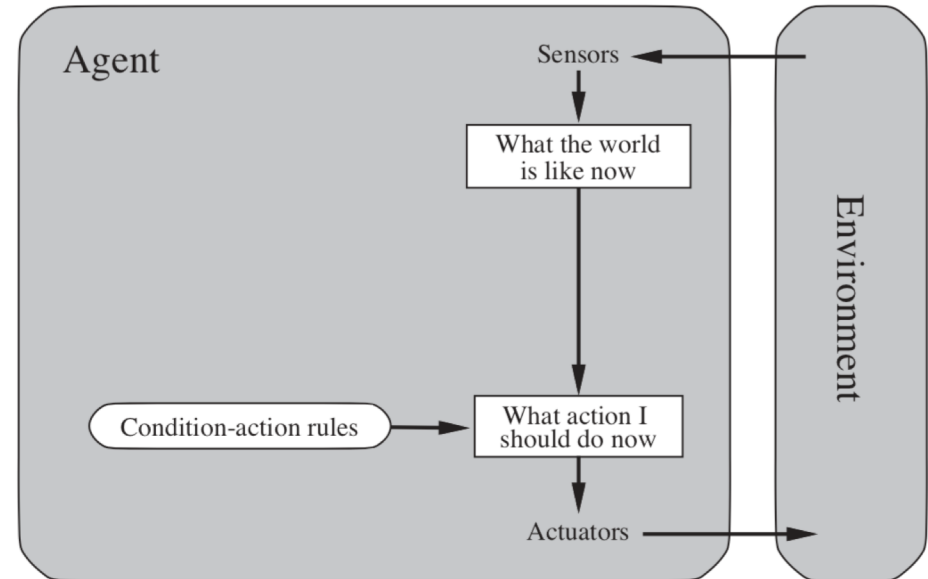
Table-driven Approach

- Implements the desired agent function
- Yet, doomed to failure
- Can AI do for general intelligent behavior?

- *The key challenge for AI is to find out how to write programs that, to the extent possible, produce rational behavior from a smallish program rather than from a vast table.*

Simple Reflex Agents

- Respond directly to percepts:
 - Choose action based on current percept (and maybe memory)
 - May have memory or a model of the world's current state
 - Do not consider the future consequences of their actions
 - Consider how the world IS
- Can a reflex agent be rational?



function SIMPLE-REFLEX-AGENT(*percept*) **returns** an action
persistent: *rules*, a set of condition–action rules

state ← INTERPRET-INPUT(*percept*)

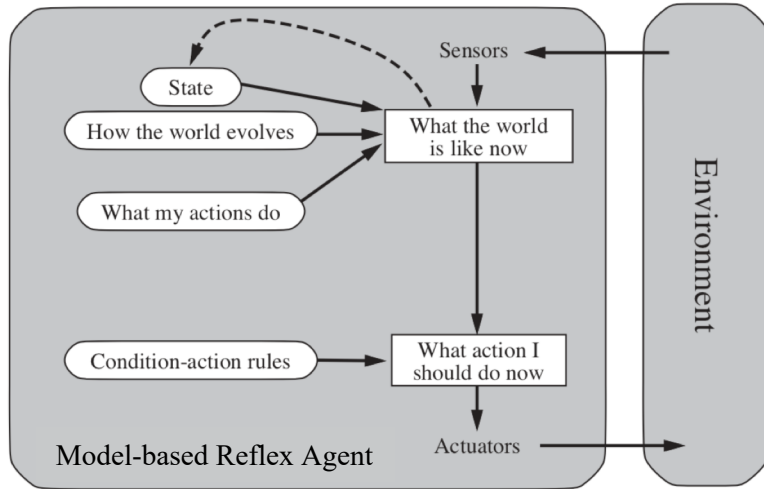
rule ← RULE-MATCH(*state*, *rules*)

action ← *rule*.ACTION

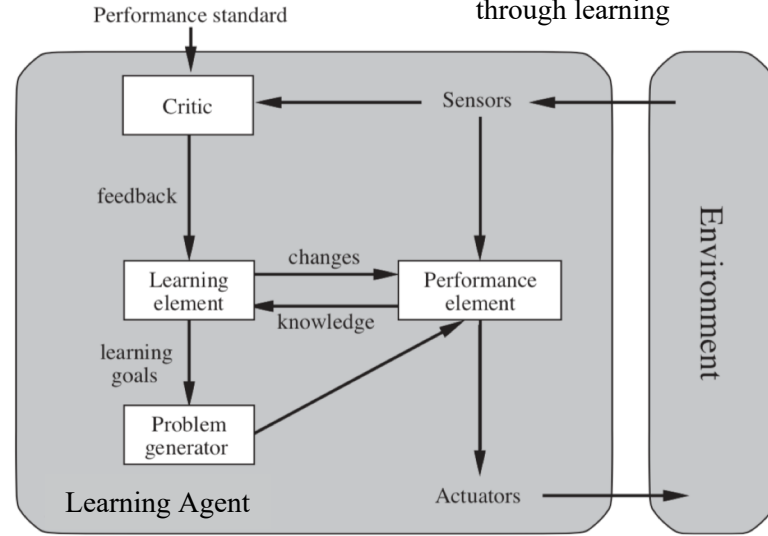
return *action*

Other Agents

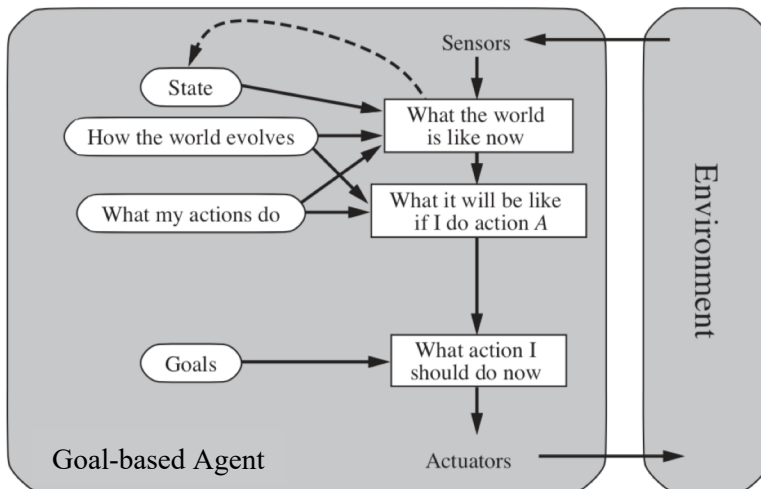
Maintain internal state to track aspects of the world that are not evident in the current percept



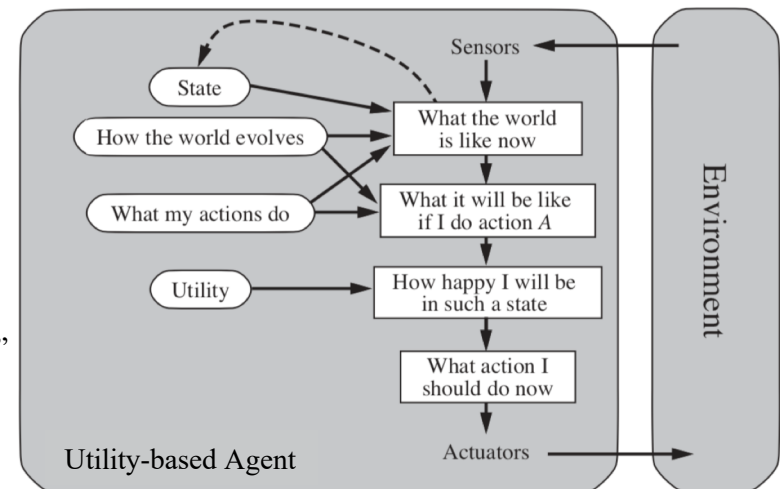
All agents can improve their performance through learning



Act to achieve their goals



Try to maximize their own expected "happiness."



How The Components Of Agent Programs Work

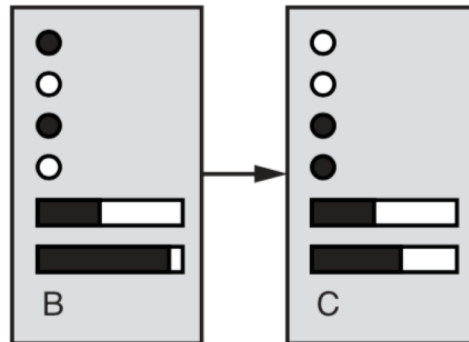
- “What is the world like now?”
 - “What action should I do now?”
 - “What do my actions do?”
- “How on earth do these components work?”

a state (such as B or C) is a black box with no internal structure



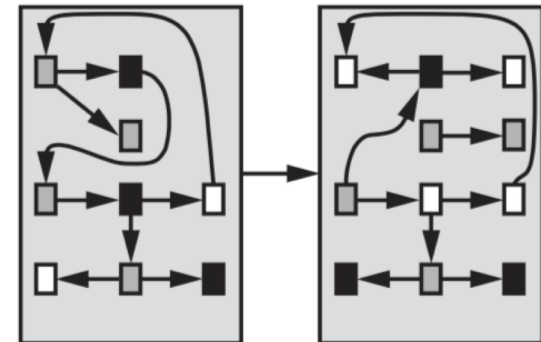
(a) Atomic

a state consists of a vector of attribute values; values can be Boolean, real-valued, or one of a fixed set of symbols



(b) Factored

a state includes objects, each of which may have attributes of its own as well as relationships to other objects



(b) Structured

Thank you!

Prof. Song Chaoyang

- Dr. Wan Fang (sophie.fwan@hotmail.com)

