인공지능

22년 삼성 AI 전문가과정
6월 9일 목요일 6교시
장병탁

21차시 : Human-Level AI

서울대학교 컴퓨터공학부
담당 교수: 장병탁

Seoul National University
Byoung-Tak Zhang

Lecture Overview

인공지능

21차시 : Human-Level AI

서울대학교 컴퓨터공학부
담당 교수: 장병탁

Seoul National University
Byoung-Tak Zhang




Introduction: Human-Level AI

❑ Where are we headed?

- What is the limit of AI?
- Can machines really think?
- The ethics of AI

❑ What remains to be done?

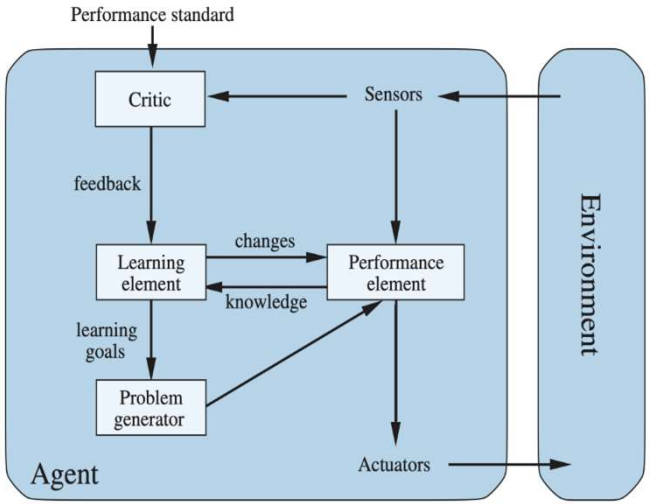
- Agent designs
- Agent components
- What's missing**?**

Philosophy, Ethics, and Safety of AI

- **The Limits of AI**
  - The argument from disability
  - The mathematical objection
  - The argument from informality
- **Can Machines Really Think?**
  - The mind-body problem
  - The Chinese room
  - Consciousness and qualia
- **The Ethics of AI**
  - What if the effects of AI are more negative than positive?
  - How to design a friendly AI?

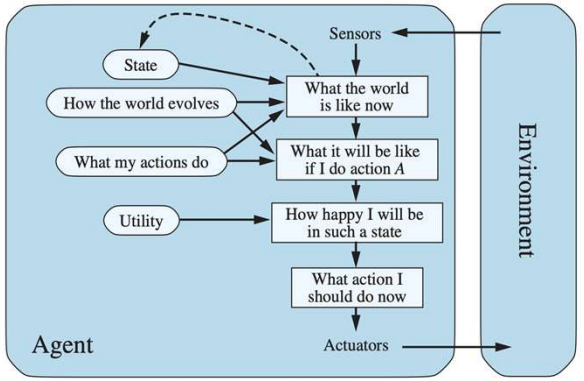
The Present of AI: What's Known

- In this course, we viewed AI as the task of **designing rational agents**. The progress we have made so far includes:
- **Agent designs**
  - Reflex, goal-based, model-based, utility-based agents
  - Knowledge-based agent, decision-theoretic agent
  - Reinforcement learning agent (POMDP agents)
- **Component technologies**
  - Reasoning: Logical, probabilistic, neural reasoning
  - Representation: Atomic, factored, structured representations
  - Learning: Various algorithms learning from various data
  - Sensors and actuators: Fully-observable, partially-observable, deterministic, stochastic



The Future of AI: What's Missing

- To achieve a **human-level general-purpose intelligent agent** that can perform well in a wide variety of environments, what's missing in components and overall architecture of an intelligent agent?
- **Agent components**
  1) Interaction with the environment through sensors and actuators
  2) Keeping track of the state of the world over time
  3) Projecting, evaluating, and selecting future courses of action
  4) Utility as an expression of preferences
  5) Learning new abstract representations gradually



- **Agent architecture**
  1) Hybrid architecture: Reflex response as well as knowledge-based deliberation
  2) Anytime algorithm: An algorithm whose output quality improves gradually over time
  3) Meta-reasoning: Decision-theoretic evaluation of value as well as cost, reflective agent

Outline (Lecture 21)

21.1 The Limits of AI

21.1 The Limits of AI (1/5)

1) Weak AI

- Weak AI hypothesis: Machines could act *as if* they were intelligent
  - Strong AI hypothesis: Machines that do so are *actually* thinking (not just *simulating* thinking)
- Weak AI is possible: "Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it."
  (McCarthy et al., 1955)

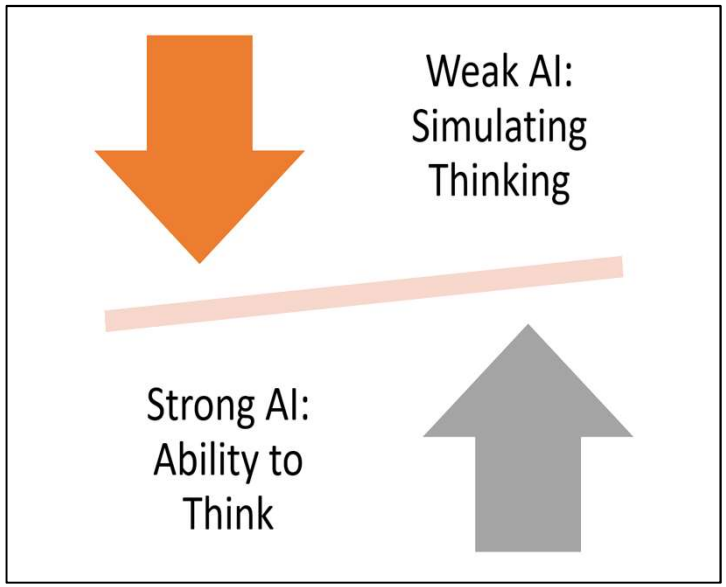
Weak AI	Strong AI
Weak AI is simply the view that intelligent behavior can be modeled and used by computers to solve complex problems.	Strong AI refers to a hypothetical machine that exhibits human cognitive abilities.
Weak AI refers to systems that are programmed to accomplish a wide range problems but operate within a pre-defined range of functions.	Strong AI refers to machines with the mind of their own and which can think and accomplish complex tasks on their own.
Weak AI-powered machines do not have mind of their own.	Strong AI-powered machines can exhibit strong human cognitive abilities.
Alexa and Siri are the best examples of weak AI programs.	Strong AI is a hypothetical concept which does not exist yet in its true form.

21.1 The Limits of AI (2/5)

1) Weak AI

- "Can machines think?": "The question of whether *Machines Can Think* . . . is about as relevant as the question of whether *Submarines Can Swim*." (Dijkstra, 1984)

- Turing Test (Alan Turing, 1950): Instead of asking whether machines can think, we should ask whether machines can pass a behavioral intelligence test, which has come to be called the Turing Test.



https://jrodthoughts.medium.com/g%C3%B6del-consciousness-and-the-weak-vs-strong-ai-debate-31afea7e0a36

21.1 The Limits of AI (3/5)

2) The argument from disability

- The claim that "a machine can never do *X*." As examples of *X*, Turing lists the following:
  - Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humor, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make someone fall in love with it
- Given what we now know about computers, it is not surprising that they do well at combinatorial problems such as playing chess.
- But algorithms also perform at human levels on tasks that seemingly involve human judgment, or as Turing put it, "*learning from experience*" and the ability to "*tell right from wrong*."

21.1 The Limits of AI (4/6)

3) The mathematical objection

- Certain mathematical questions are in principle unanswerable by particular formal systems, e.g. Gödel's incompleteness theorem.
  - For any formal axiomatic system F powerful enough to do arithmetic, it is possible to construct a so-called Gödel sentence $G(F)$ with the following properties:
    - $G(F)$ is a sentence of F but cannot be proved within F.
    - If F is consistent, then $G(F)$ is true.

$$G \leftrightarrow \neg \mathrm{Prov}(\ulcorner G \urcorner)$$

https://www.cantorsparadise.com/g%C3%B6dels-first-incompleteness-theorem-in-simple-symbols-and-simple-terms-7d7020c28ac4

21.1 The Limits of AI (5/6)

4) The argument from informality

- This is the claim that the human behavior is far too complex to be captured by any simple set of rules and that because computers can do no more than follow a set of rule, they cannot generate behavior as intelligent as that of humans.
- The inability to capture everything in a set of logical rules is called the *qualification problem in AI*.
- The position they criticize came to be called "*Good Old-Fashioned AI*," or GOFAI, a term coined by philosopher John Haugeland (1985).

21.1 The Limits of AI (6/6)

4) The argument from informality (cont.)

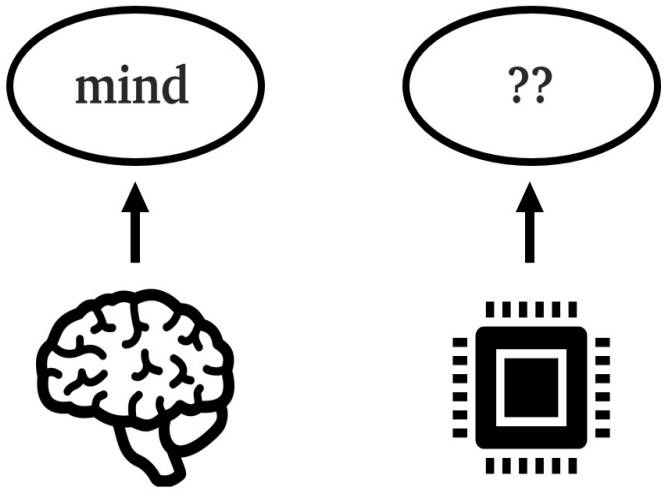
- Dreyfus and Dreyfus (1986)'s *Mind over Machines*: points out several problems of AI, but these have been addressed with partial success and some with total success:
  - Good generalization from examples cannot be achieved without background knowledge.
  - It cannot operate autonomously without the help of a human trainer.
  - Learning algorithms do not perform well with many features, and if we pick a subset of features, "there is no known way of adding new features should the current set prove inadequate to account for the learned facts."

21.2 Can Machines Really Think?

21.2 Can Machines Really Think? (1/7)

1) Strong AI

- Strong AI hypothesis: Machines that think are *actually* thinking
  (not just *simulating* thinking)
  - (cf.) Weak AI hypothesis: Machines could act *as if* they were intelligent
- The objection: Many philosophers have claimed that a machine that passes the *Turing Test* would still not be actually thinking, but would be only a simulation of thinking.
  - **Consciousness, Phenomenology, Intentionality**



21.2 Can Machines Really Think? (2/7)

1) Strong AI (cont.)

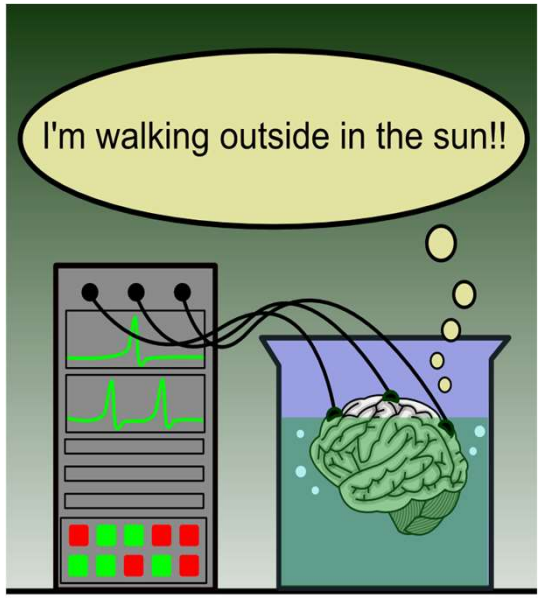
- Mind-body problem: The question of whether machines could have real minds is directly relevant to the philosophical efforts to solve the mind-body problem:
  - Do humans have real mind? / Dualist / Monist

- The problem for physicalists is to explain how physical states—in particular, the molecular configurations and electrochemical processes of the brain—can simultaneously be mental states, such as being in pain, enjoying a hamburger

21.2 Can Machines Really Think? (3/7)

2) Mental states and the brain in a vat

- If physicalism is correct, it must be the case that the proper description of a person's mental state is *determined* by that person's brain state.
- The simplicity of this view is challenged by some simple thought experiments "brain in a vat".
- This example seems to contradict the view that brain states determine mental states.
  - It would be literally false to say that you have the mental state "knowing that one is eating a hamburger".

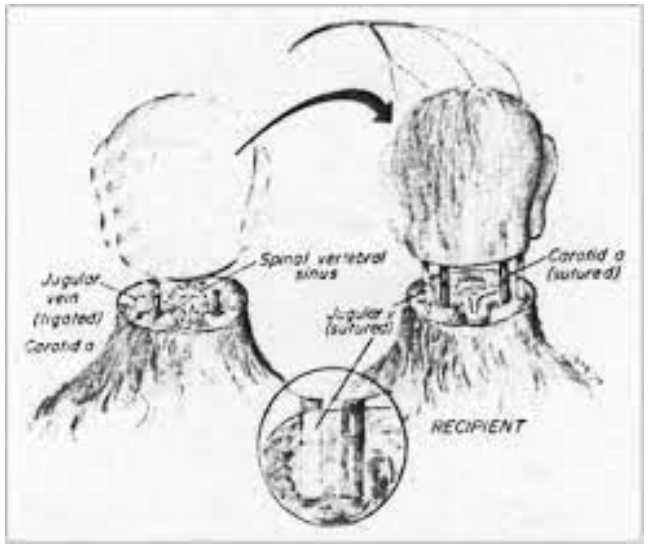


Brain in a VAT. Psychology Wiki. (n.d.). Retrieved February 25, 2022, from https://psychology.fandom.com/wiki/Brain_in_a_vat

21.2 Can Machines Really Think? (4/7)

3) Functionalism and the brain replacement experiment

- Functionalism: Under functionalist theory, any two systems with isomorphic causal processes would have the same mental states.
  - Therefore, a computer program could have the same mental states as a person.
- The claims of functionalism are illustrated most clearly by the brain replacement experiment

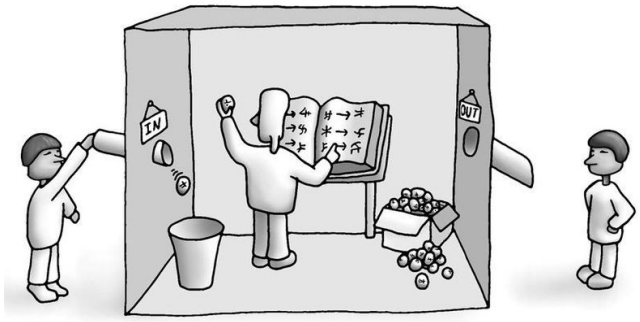


https://www.google.com/search?q=brain+replacement+experiment&tbm=isch&ved=2ahUKEwjjzOvjpKz3AhXGUPUHHXKZAhYQ2-cCegQIABAA&oq=brain+replacement+experiment&gs_lcp=CgNpbWcQAzoFCAAQgAQ6BggAEAcQHlD7A1j7A2DlBWgAcAB4AIABpwGIAZsCkgEDMC4ymAEAoAEBqgELZ3dzLXdpei1pbWfAAQE&sclient=img&ei=FwtlYqPlDMah1e8P8rKKsAE&bih=789&biw=1394#imgrc=oh3F93h9kmiM_M

21.2 Can Machines Really Think? (5/7)

4) Biological naturalism and the Chinese Room

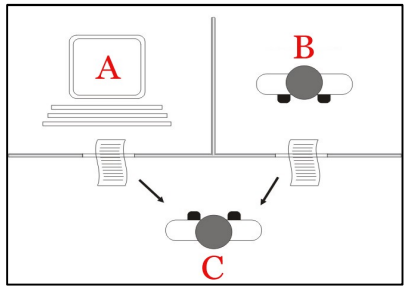
- A strong challenge to functionalism has been mounted by John Searle's (1980) biological naturalism: "Mental states are high-level emergent features that are caused by low-level physical processes *in the neurons*"
- Thus, the mental states cannot be duplicated just on the basis of some program having the same functional structure with the same input-output behavior
- To support this view, he proposed the Chinese Room thought experiment.



21.2 Can Machines Really Think? (6/7)

Chinese Room

- Human in a room doesn't know Chinese, but she has a rulebook for translating Chinese letters
- If she is good at this translation, an observer outside the room will think she is fluent at Chinese
- She never understood Chinese, but just followed the rulebook.
- Conclusion (by Searle)
  - Understanding is not necessary to solve the problem
  - Computer programs: Syntactic
  - Human minds: Semantic
  - Syntax by itself is neither constitutive of nor sufficient for semantics



Source : https://commons.wikimedia.org/wiki/File:Test_de_Turing.jpg

21.2 Can Machines Really Think? (7/7)

5) Consciousness, qualia, and the explanatory gap

- Running through all the debates about strong AI—the elephant in the debating room — is the issue of *consciousness*.
- Aspects of consciousness: Understanding, self-awareness, subjective experience (feels like something to have certain brain states)
- Qualia: intrinsic nature of experiences
- Explanatory gap: neuroscience to cognitive science
  - Behaviors of several neurons can't explain whole process of cognitive behavior
  - Humans are simply incapable of forming a proper understanding of their own consciousness

21.3 The Ethics of AI




21.3 The Ethics of AI (1/3)

So far, we have concentrated on whether we *can* develop AI, but we must also consider whether we *should*.

1) Risks of AI

- What if the effects of AI technology are more likely to be negative than positive? In fact, AI poses some fresh problems, such as:
  - People might lose their jobs to automation.
  - People might have too much (or too little) leisure time.
  - People might lose their sense of being unique.
  - AI systems might be used toward undesirable ends.

21.3 The Ethics of AI (2/3)

2) Three sources of bigger risks of AI

- **State estimation may be incorrect, causing agent to do the wrong thing**
  - Humans make more mistakes
  - Design a system with checks and balances
- **Finding right utility function is not easy**
  - Reducing human suffering: no human, no suffering?
  - Techniques, such as apprenticeship learning, allow us to specify utility functions
- **Learning function may cause agent to evolve into unintended behaviors**
  - Ultra-intelligent machines (Good, 1965)
  - Intelligence explosion, technological singularity

21.3 The Ethics of AI (2/3)

3) Three laws of robotics (Asimov, 1942)

- A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- A robot must obey orders given to it by human beings, except where such orders would conflict with the First Law.
- A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

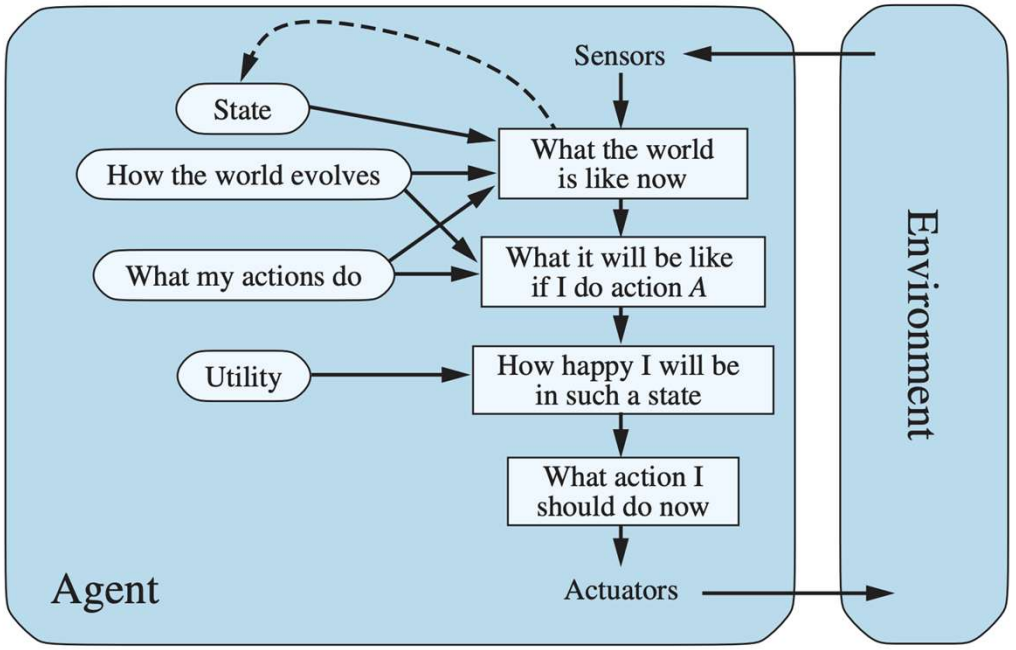
4) How to design a Friendly AI?

- Friendliness (a desire not to harm humans) should be designed in from the start, but the designers should recognize that the robot will learn and evolve over time.
- To define the mechanism for evolving AI systems under a system of checks and balances.

21.4 AI Components

21.4 AI Components (1/3)

- Let's look at the components of an intelligent agent to assess what's known and what's missing. We consider the utility-based agent with a learning component and see where the state-of-the-art stands for each of the components.

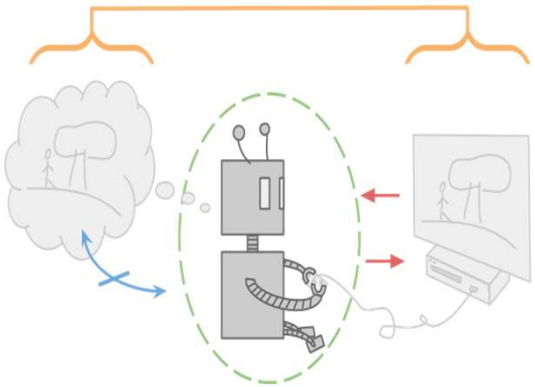


<출처> Stuart J. Russell and Peter Norvig (2021). Artificial Intelligence: A Modern Approach (4th Edition). Pearson

21.4 AI Components (2/3)

Advances and opportunities for further progress

- **1) Interaction with the environment through sensors and actuators**
  - Present: Availability of ready-made programmable robots, and sensors and actuators getting more elaborate
  - Future: AI systems are at the cusp of moving from software-only systems to embedded robotic systems
- **2) Keeping track of the state of the world**
  - Present: Filtering algorithms for probabilistic reasoning in atomic and factored state representations
  - Future: Probability and first-order logic representations coupled with aggressive machine learning



21.4 AI Components (3/3)

Advances and opportunities for further progress

- **3) Projecting, evaluating, and selecting future courses of action**
  - Present: Hierarchical reinforcement learning has succeeded for decision making
  - Future: How the search for effective long-range plans might be controlled
- **4) Utility as an expression of preferences**
  - Present: Rational decisions based of maximization of expected utility.
  - Future: Knowledge engineering for reward functions to convey to the agents what we want them to do
- **5) Learning**
  - Present: Machine learning today assumes a factored representation
  - Future: Gradually constructing new representations at levels of abstraction higher than the input vocabulary

21.5 AI Architectures




21.5 AI Architectures (1/5)

1) Hybrid architecture

- Reflex responses are needed for situations in which time is of the essence, whereas knowledge-based deliberation is needed for plan ahead.
- A complete agent must be able to do both, using a hybrid architecture.
- Agents also need ways to control their own deliberations. Cease deliberating when action is demanded and use the time available for deliberation to execute the most profitable computations.
- Real-time deliberation is also important: real-time AI

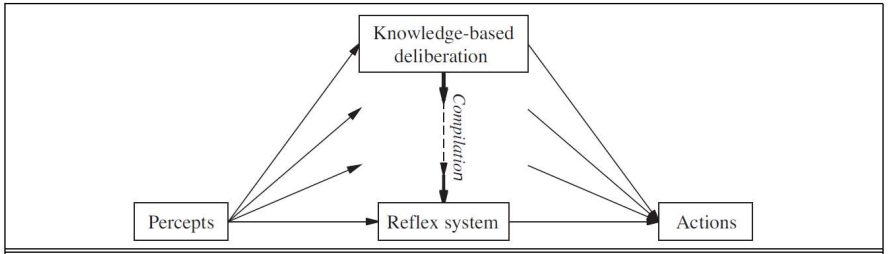


Figure 27.2 Compilation serves to convert deliberative decision making into more efficient, reflexive mechanisms.

<출처> Stuart J. Russell and Peter Norvig (2021). Artificial Intelligence: A Modern Approach (4th Edition). Pearson

21.5 AI Architectures (2/5)

2) General methods of controlling deliberation

- **1**. Employ anytime algorithms
  - An algorithm whose output quality improves gradually over time.

- 2. Decision-theoretic metareasoning
  - The value of a computation depends also on its cost
  - Metareasoning can be used to design better search algorithms and to guarantee that the algorithms have the anytime property.

- 3. Reflective architecture: Meta-reasoning is one specific example of a reflective architecture.

21.5 AI Architectures (3/5)

Whether AI's current path is more like a tree climb or a rocket trip?

1) Rationally acting agents: four possibilities

- Perfect rationality: finds best way to maximize its own expected utility always, but it is too time consuming (not realistic)
- Calculative rationality: a calculative rational agent eventually returns what would have been the rational choice at the beginning of its deliberation
- Bounded rationality: deliberating only long enough to come up with an answer that is "good enough" (or satisficing) (Simon, 1957)
- Bounded optimality: a bounded optimal agent behaves as well as possible, given its computational resources

21.5 AI Architectures (4/5)

2) Bounded optimality

- Bounded optimal (BO) agents are actually useful in the real world
  - Calculative rationality (design) → Bounded optimality (implement)
- Yet, no idea what BO programs are like for large, general-purpose computers in complex environments.
- Asymptotic bounded optimality (ABO)
  - Relaxed version of bounded optimality
  - Suppose a program P is BO for a machine M in a class of environments E,
  - Then program P' is ABO for M in E if it can outperform P by running on a machine kM that is k times faster (or larger) than M.
- Unless k were enormous, we would be happy with a program that was ABO for a nontrivial environment on a nontrivial architecture.

21.5 AI Architectures (4/5)

Will AI be used for good or ill?

- AI developers have a responsibility to see that the impact of their work is a positive one. The scope of impact will depend on the degree of successes of AI.
- Even moderate successes in AI have already changed the ways in which computer science is taught and software development is practiced.
- Medium-level successes in AI would affect all kinds of people in their daily lives.
  - Personal assistants, automated driving, autonomous weapons, genomics,
- Large-scale success in AI—the creation of human-level intelligence and beyond —will change our life and future of human race.
- Alan Turing(1950): *We can see only a short distance ahead, but we can see that much remains to be done.*

Summary

- We use the term weak AI for the hypothesis that machines could possibly behave intelligently, and strong AI for the hypothesis that such machines would count as having actual minds (as opposed to simulated minds).
- Alan Turing rejected the question "Can machines think?" and replaced it with a behavioral test. He anticipated many objections to the possibility of thinking machines.
- We identified six potential threats to society posed by AI and related technology. One threat in particular is worthy of further consideration: that ultraintelligent machines might lead to a future that is very different from today.
- Components of an intelligent agent to assess what's known and what's missing: Interaction with the environment, keeping track of the state of the world, projecting/evaluating/selecting future courses of action, utility, and learning.
- For controlling deliberation, employ **anytime algorithms**, or apply **decision-theoretic metareasoning**.
- The goal of AI: Perfect rationality, calculative rationality, bounded rationality, bounded optimality.
- We can expect that medium-level successes in AI would affect all kinds of people in their daily lives.
- It seems likely that a large-scale success in AI—the creation of human-level intelligence and beyond—would change the lives of a majority of humankind.