인공지능

23차시 : Human-Level AI

서울대학교 컴퓨터공학부
담당 교수: 장병탁
Seoul National University
Byoung-Tak Zhang

Introduction

- **Philosophical Foundations of AI: What it means to think and whether artifacts could and should ever do so?**
  - How do minds work?
  - Is it possible for machines to act intelligently in the way that people do, and if they did, would they have real, conscious minds?
  - What are the ethical implications of intelligent machines?
- **The Present and Future of AI: Where we are and where we are going, this being a good thing to do before continuing?**
  - Will all this progress lead to a 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?
  - Whether designing rational agents is the right goal in the first place?
  - What if AI does succeed? Consequences of success in our endeavor

Outline (Lecture 23)

Stuart Russell & Peter Norvig, Artificial Intelligence: A Modern Approach (3rd Edition), Chapters 26 & 27

23.1 Weak AI: Can Machines Act Intelligently? (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)
- "**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.
- **Possible objections to the possibility of intelligent machines**: (see next pages)

23.1 Weak AI: Can Machines Act Intelligently? (2/5)

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, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behavior as man, do something really new.
- 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*."

23.1 Weak AI: Can Machines Act Intelligently? (3/5)

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.
- Philosopher Lucas have claimed that this theorem shows that machines are inferior to humans, because machines are formal systems that are limited by the incompleteness theorem while humans have no such limitation. Three of the problems with this claim:
  - First, Gödel's incompleteness theorem applies only to formal systems that are powerful enough to do arithmetic.
  - Second, an agent should not be too ashamed that it cannot establish the truth of some sentence while other agents can.
  - Third, and most important, even if we grant that computers have limitations on what they can prove, there is no evidence that humans are immune from those limitations.

23.1 Weak AI: Can Machines Act Intelligently? (4/5)

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**.
- Under Dreyfus's view, human expertise does include knowledge of some rules,
  - but only as a "holistic context" or "background" within which humans operate.
- The position they criticize came to be called "Good Old-Fashioned AI," or **GOFAI**, a term coined by philosopher John Haugeland (1985).
- **GOFAI** is supposed to claim that all intelligent behavior can be captured by a system that reasons logically from a set of facts and rules describing the domain.

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.
    - In fact, we saw in Chapters 19 and 20 that there are techniques for using prior knowledge in learning algorithms.
  - It cannot operate autonomously without the help of a human trainer.
    - In fact, learning without a teacher can be accomplished by **unsupervised learning** (Chapter 20) and **reinforcement learning** (Chapter 21).
  - 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."
    - In fact, new methods such as support vector machines handle large feature sets very well. Recent deep learning methods can learn by generating new features.
  - "Currently, no details of this mechanism are understood or even hypothesized in a way that could guide AI research."
    - In fact, the field of active vision, underpinned by the theory of information value (Chapter 16), is concerned with exactly the problem of directing sensors, and already some robots have incorporated the theoretical results obtained.

23.2 Strong AI: 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.
  - The objection by Professor Geoffrey Jefferson (1949): Not until a machine could write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain—that is, not only write it but know that it had written it.
- **Consciousness**: Turing calls this the argument from consciousness—the machine has to be aware of its own mental states and actions.
- **Phenomenology**: While consciousness is an important subject, Jefferson's key point actually relates to phenomenology, or the study of direct experience: the machine has to actually feel emotions.
- **Intentionality**: Others focus on **intentionality**—that is, the question of whether the machine's purported beliefs, desires, and other representations are actually "about" something in the real world.

23.2 Strong AI: 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?
    - Real mind (like spirit) vs. neurophysiological process
  - Dualist: Descartes, *Meditations on First Philosophy*
    - Mind can be separated from body, mind-body dualism
  - Monist: Physicalism
    - Mind can't be separate from the body: Mental states = physical states
- Most modern philosophers of mind are physicalists of one form or another, and physicalism allows, at least in principle, for the possibility of strong AI.
- 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, knowing that one is riding a horse, or believing that Vienna is the capital of Austria.

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".
  - *Imagine that your brain was removed from your body at birth and placed in a marvelously engineered vat. The vat sustains your brain, allowing it to grow and develop. Electronic signals are fed to your brain from a computer simulation of an entirely fictitious world, and motor signals from your brain are intercepted and used to modify the simulation as appropriate. In fact, the simulated life you live replicate exactly the life you would have lived, had your brain not been placed in the 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".
- **Resolving the dilemma is to say that the content of the mental states can be interpreted from two different points of view:**

1. **Wide content:** Omniscient outside observer with access to the whole situation. Both the brain state and environment history. Relevant if one's goals are to ascribe mental states to others who share one's world, to predict their behavior and its effects.
2. **Narrow content:** Only the brain state. The narrow content of the brain states of a real hamburger-eater and a brain-in-a vat "hamburger"-"eater" is the same in both cases. Relevant if one is concerned with the question of whether AI systems are really thinking.

23.2 Strong AI: 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
  - *Suppose that some miraculous surgical technique can replace individual neurons with the corresponding electronic devices without interrupting the operation of the brain as a whole. The experiment consists of gradually replacing all the neurons in someone's head with electronic devices.*
- We are concerned with both the external behavior and the internal experience of the subject, during and after the operation. We must have an explanation of the manifestations of consciousness produced by the electronic brain that appeals only to the functional properties of the neurons. *And this explanation must also apply to the real brain, which has the same functional properties.* There are three possible conclusions:
  - The causal mechanisms of consciousness that generate these kinds of outputs in normal brains are still operating in the electronic version, which is therefore conscious.
  - The conscious mental events in the normal brain have no causal connection to behavior, and are missing from the electronic brain, which is therefore not conscious.
  - The experiment is impossible, and therefore speculation about it is meaningless.

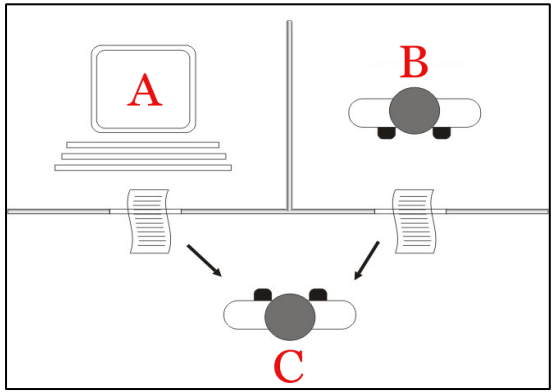
4) Biological naturalism and the Chinese Room

- 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
  - We would require that the program be running on an architecture with the same causal power as neurons.
- To support this view, he proposed the Chinese Room thought experiment.
  - Searle's claim rests upon the following four axioms (Searle, 1990):
    1. Computer programs are formal (syntactic).
    2. Human minds have mental contents (semantics).
    3. Syntax by itself is neither constitutive of nor sufficient for semantics.
    4. Brains cause minds.

23.2 Strong AI: 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.



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

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

- Artificial brain would have to duplicate the causal powers of brains, not just run a particular program.
- Even if we find a program that is mimicking human brain behaviors, that doesn't mean it is a mind.

23.2 Strong AI: 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
  - Qualia present a challenge for functionalist accounts for the mind because different qualia could be involved in what are otherwise isomorphic causal processes
  - Qualia are challenging not just for functionalism but for all of science
- **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
  - Dennett (1991): avoids the gap by denying the existence of qualia, attributing them to a philosophical confusion

23.3 The Ethics and Risks of Developing 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.
  - The use of AI systems might result in a loss of accountability.
  - The success of AI might mean the end of the human race.

23.3 The Ethics and Risks of Developing 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
  - Mind children (Moravec 2000), transhumanism (future in which humans are merged with robotic and biotech inventions) (Kurzweil, 2005)

23.3 The Ethics and Risks of Developing AI (3/3)

3) Three laws of robotics (Asimov, 1942)

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given to it by human beings, except where such orders would conflict with the First Law.
3. 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.
- To give the systems utility functions that will remain friendly in the face of such changes.

23.4 Agent Components (1/2)

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.

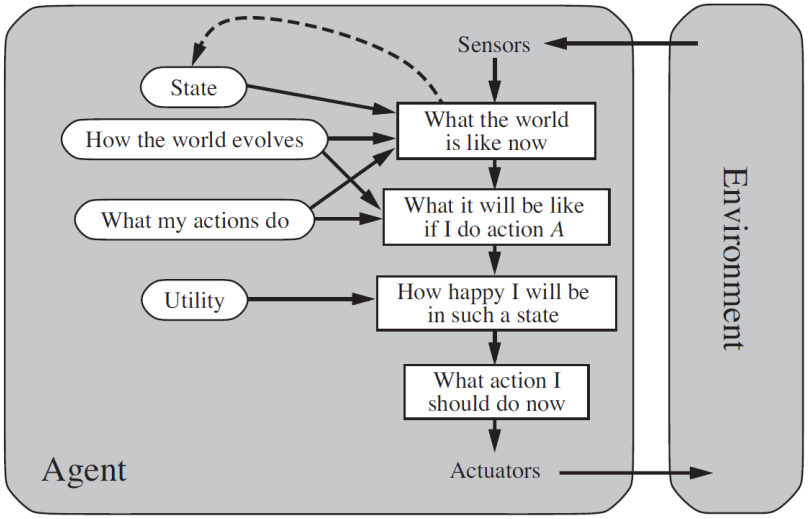


Figure 27.1 A model-based, utility-based agent, as first presented in Figure 2.14.

사진 출처 #1

23.4 Agent Components (2/2)

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

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. Little work on realistic utility functions
- 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 for inductive learning of functions
- Future: Gradually constructing new representations at levels of abstraction higher than the input vocabulary

23.5 Agent Architectures (1/2)

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

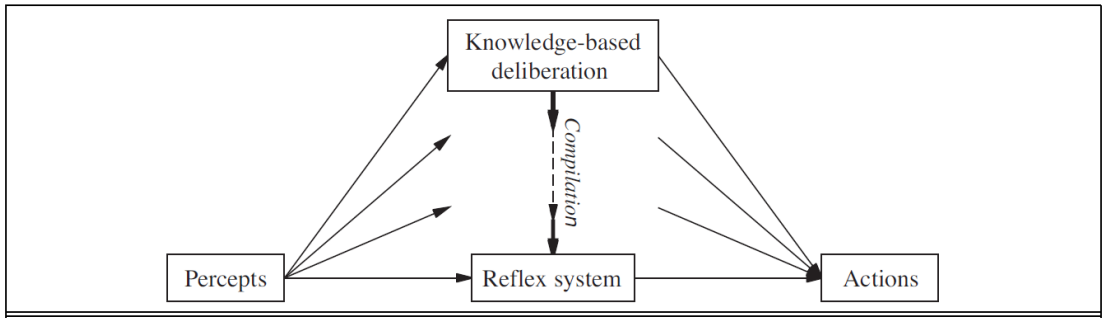


Figure 27.2 Compilation serves to convert deliberative decision making into more efficient, reflexive mechanisms.

사진 출처 #2

23.5 Agent Architectures (2/2)

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 on both its cost (in terms of delaying action) and its benefits.
  - Metareasoning can be used to design better search algorithms and to guarantee that the algorithms have the anytime property.
- **Reflective architecture:** Metareasoning is one specific example of a reflective architecture.
  - An architecture that enables deliberation about the computational entities and actions occurring within the architecture itself.
  - Decision-making and learning algorithms can be designed that operate over the joint space of the environment state and the computational state and thereby serve to implement and improve the agent's computational activities.

23.6 Are We Going in the Right Direction? (1/2)

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

23.6 Are We Going in the Right Direction? (2/2)

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**,
    - where the complexity of environments in **E** is unbounded.
  - 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.

23.7 What If AI Does Succeed?

Will AI be used for good or ill?

- There are ethical issues to consider. 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.
  - Speech recognition, inventory control systems, surveillance systems, robots, search engines, self-driving cars, etc.
- Medium-level successes in AI would affect all kinds of people in their daily lives.
  - Personal assistants, automated driving, autonomous weapons, genomics, energy management, verification of treaties concerning nuclear weapons
- Large-scale success in AI—the creation of human-level intelligence and beyond —will change our life and future of human race.
  - Nature of our work and play, our view of intelligence, consciousness, and the destiny of human race, threat to human autonomy, freedom and even survival.
- Alan Turing (1950): *We can see only a short distance ahead, but we can see that much remains to be done.*

Summary (1/2)

- Philosophers 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. Few AI researchers pay attention to the Turing Test, preferring to concentrate on their systems' performance on practical tasks, rather than the ability to imitate humans.
- We identified six potential threats to society posed by AI and related technology. We concluded that some of the threats are either unlikely or differ little from threats posed by "unintelligent" technologies. One threat in particular is worthy of further consideration: that ultraintelligent machines might lead to a future that is very different from today—we may not like it, and at that point we may not have a choice. Such considerations lead inevitably to the conclusion that we must weigh carefully, and soon, the possible consequences of AI research.

Summary (2/2)

- 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, 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.

Homework

- What are the capabilities of AI still missing to achieve a truly human-level general intelligence?

출처

사진

1, 2 Stuart J. Russell and Peter Norvig (2016). Artificial Intelligence: A Modern Approach (3rd Edition). Pearson