9. History and Philosophy of Artificial Intelligence

9.1 What is AI?

- Study of systems that act in a way that looks intelligent to any observer.
  - Attempt to understand human intelligence by replicating its functionality.
- Application of methods based on intelligent behavior of humans or other animals to solve complex problems, for instance
  - Global warming
  - Natural disaster rescue and prediction
  - Wars and other human conflicts
  - Medical diagnosis
- Build systems to aid human user.

9.2 First milestones

- Neural network model with binary neurons by Warren McCulloch and Walter Pitts in 1943.
  - Equivalent to the Turing machine
  - It could learn
  - Limited: neurons have highly non-linear characteristics
- LISP by John McCarthy in 1950's.
- General Problem Solver (GPS) by Allen Newell and Herbert Simon in 1960's.
  - Means-ends analysis
  - Based on formal logic
  - Failed to solve complicated problems

9.3 AI in the 1960's

- General methods of solving broad classes of problems
- Toy problems
- No knowledge
- Clever search algorithms
- First attempt for machine translation failed ← requires general understanding of the subject

9.4 AI in the 1970's

- Shift from general purpose, knowledge-sparse, weak methods to domain specific, knowledge-intensive strong methods
- Problem domains need to be restricted
- Expert systems
- DENDRAL (Buchanan et al., 1969) to analyze chemicals:
  - NASA project
  - Extract knowledge from human experts and map it to highly specific rules
  - Knowledge engineering
  - Commercially marketed in the US

- MYCIN (Feigenbaum et al., 1972, Shortliffe, 1976)
  - diagnosis of infectious blood diseases
  - Provided therapeutic advice
  - Knowledge consisted of 450 independent IF-THEN rules
  - Uncertain reasoning involved
  - No knowledge of human physiology
  - Domain-independent version EMYCIN (1979)

- Great success of expert systems reported in several fields: chemistry, electronics, engineering, geology, medicine, process control, and military science.
- However,
  - Narrow domain expertise
  - Not as robust and flexible as preferred.
  - Often fail in atypical problems in unpredictable ways.
  - No deep understanding how knowledge relates to domain.
  - Hard to verify and validate, or identify incomplete, incorrect or inconsistent knowledge.
  - Little ability to learn from experience

9.5 AI in 1980's and thereafter

- Revival of the neural network research.
  - Hopfield networks (Hopfield, 1982)
  - Self-organizing maps (Kohonen, 1982)
  - Reinforcement learning (Barto, Sutton, & Anderson, 1983)
  - PDP (Rumelhart & McClellan, 1986)
  - Advances in computer technology.
- Evolutionary computing
  - The idea of genetic algorithms by Holland in 1975
  - Schema theorem by Goldberg, 1989.
  - Genetic programming by Koza in 1990's: genetic operator to evolve LISP code.
- Fuzzy logic
- Biologically inspired methods
  - SWARM intelligence
  - ABM

9.6 What is philosophy?

- 'Love of wisdom'
  - From Greek words 'philo' (= to love or befriend), and 'sophia' (= to be wise).
- Investigates (Sloman, 1985)
  - The most general questions about what exists;
  - The most general questions about questions and possible answers;
  - The most general questions about what ought to exist, or ought not to exist.

Ontology

Epistemology

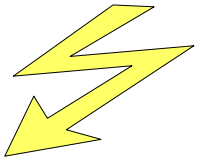
Ethics

9.7 Philosophy of mind

- Branch of philosophy that studies mind, mental functions, mental properties, consciousness, and the nature of their relationship to physical body.
- Central question is so called *mind-body problem*.

Dualism

Separate existence of mind and body.

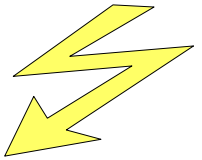


Monism

There is only one substance.

Reductive

All mental states and properties will be explained by neuroscientific accounts of brain processes and states.



Non-reductive

Mental descriptions cannot be reduced to the language and lower-level explanations of physical science.

9.8 Mind-body problem

- Dualist solutions
  - Mind and mental phenomena are non-physical.
  - Plato and Aristotle alleged that man's intelligence cannot be explained in terms of their physical body.
  - René Descartes (1596-1650) was first to associate mind with consciousness and self-awareness, and distinguish it from brain.
  - *Cartesian (interaction) dualism*: the material body and immaterial mind, while being distinct, do interact.
  - Epiphenomenalism
    - Only physical events have effect — they can cause other physical events and mental events ~~-~~but mental events are inefficacious.

- Monist solutions
  - *Physicalism* or *materialism*
    - The only existing substance is physical.
    - Mental can be reduced to physical.
  - *Idealistic monism, idealism, phenomenalism*
    - Bertrand Russell
    - The only existing substance is mental.
    - All that exists is just representations of external objects in our senses, not the objects themselves.
  - *Behaviorism*
    - Behavior can be studied and understood without referring to internal mental states.
    - Reaction to *introspectionism*.

- *Type identity theory*
  - Postulates that mental states equal brain states.
  - Do creatures having the same kind of mental states — say feeling of pain — have the same brain states?
- *Functionalism*
  - Hilary Putnam, Jerry Fodor
  - Mental states (beliefs, desires) characterized by their causal relation with other mental states and with sensation and action.
  - Multiply realizable computers
- *Eliminative materialism*
  - Paul and Patricia Churchland
  - Mental states are fictitious entities introduced by folk-psychology:
    - Common-sense understanding is false; some classes of mental states people believe in do not exist, e.g., beliefs and desires.
    - Behavior can only be explained on biological level.

9.9 Consciousness

- In simplified terms, awareness of one's existence and thought processes.
- People are not aware of all or most of their mental processes.
  - People do not plan exactly what they are going to say.
  - Blindsight – damage in the visual cortex.
- Does one have to be conscious in order think or understand?
- Different levels of consciousness and different levels of understanding.
- If every cell in the brain is replaced one by one with an equivalent computational device, would the result be intelligent? Conscious?

9.10 Belief attribution

- What cognitive properties or beliefs to attribute to non-human animals, infants, computers or robots?
- Holding a belief ≡ the brain is in particular state.
- Two views that are seen mutually exclusive and exhaustive.
- *Realism*:

  A person has a particular belief.
  A person is infected with a particular virus.

  } Equivalent questions

  - Perfectly objective criterion.
- *Interpretationism*

  A person has a particular belief.
  A person is immoral.
  A person has style or talent.
  A person makes a good spouse.

  } Equivalent questions

  - Well, it depends ...

9.11 Predicting behavior

- How to understand or predict the behavior of a system?
- Try to discern the beliefs the system has.
- Even if a belief is objective phenomenon
  - It can be only detected if one has a certain predictive strategy.
  - It can be confirmed by the assessment of success of the strategy.
- Which strategy (stance) to adopt?

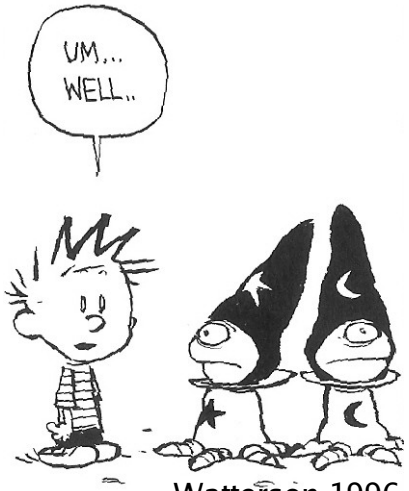
- Daniel C. Dennett (1981, 1987)
- Astrological stance
  - Feed in the date and time of birth, and come up with a perfect prediction of person's future behavior.
  - Predictions are sheer luck, vague or ambiguous to so that anything can confirm them.
  - However, good strategy not only works, but we know why it works.
- Physical stance
  - Determine the system's physical constitution, and use the knowledge on physics to predict the outcome with any input.
  - Works in principle.

- Design stance
  - Ignore the physical constitution, and assume that the system behaves as it is designed to behave under various circumstances.
  - For instance, computer, alarm clock or biological objects such as plants and animals, kidneys and hearts.
  - Obviously, only designed behavior can be predicted.
- Intentional stance
  - Treat the system as a rational agent

    → figure out the beliefs the agent ought to have

    → figure out the desires the system ought to have

    → predict that the rational agent will act to achieve its goals given the beliefs.



Watterson 1996

9.12 Intentional stance

Intentionality here means 'aboutness'; the property of mind of being directed or about objects and events in the world

- How to assign beliefs and desires (preferences, goals, interests) to people?
- Beliefs that system ought to have:
  - If some one is exposed to something, one comes to know about it.
  - All the truths relevant to system's interests.
  - Arcane and sophisticated beliefs.
- Desires the system ought to have:
  - Basic desires: for survival, absence of pain, food, comfort, entertainment.

- *"I want a two-egg mushroom omelet, some French bread and butter, and a half a bottle of lightly chilled white Burgundy."*
- Language enables us to formulate highly specific desires, but it also forces us to commit to more stringent conditions that we want to (or think of):
  - *"I'd like to have some baked beans, please."*
  - *Yes, sir. How many?"*
- Intentional strategy not only works with human beings, but with other animals and even chess-playing computers or thermostat.

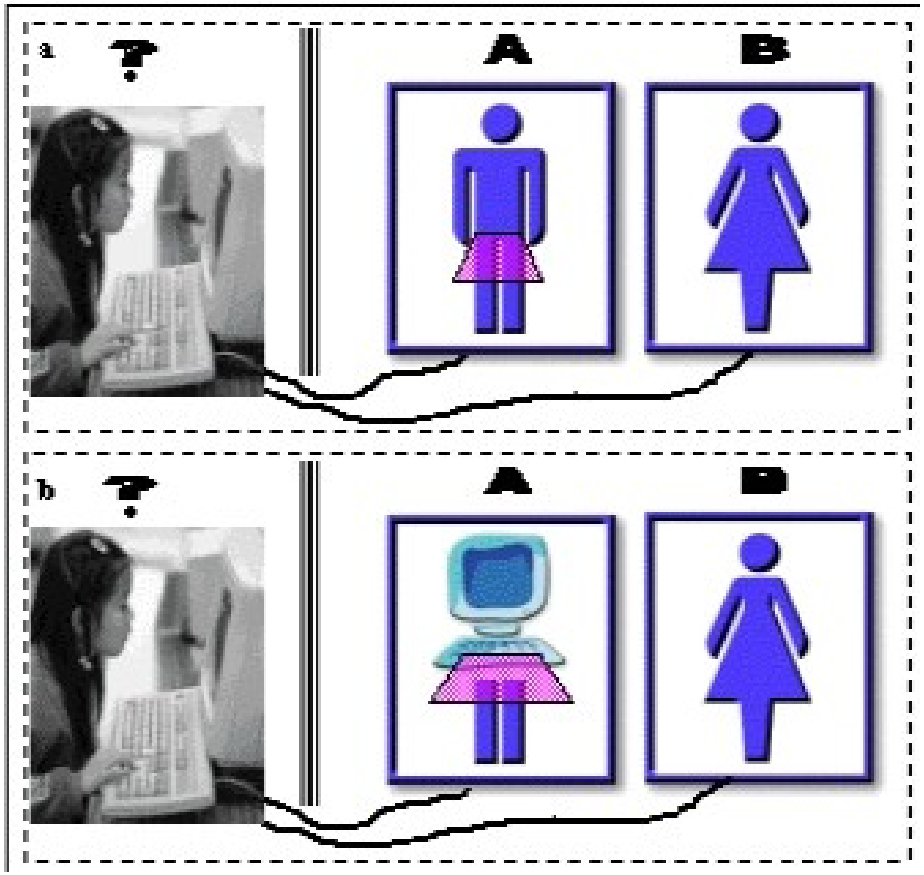
9.13 Strong vs. weak AI

- *Weak AI*
  - Human intelligence can be simulated, but the machine doing the simulation need not have conscious thought or be thinking.
  - No claim is made about the ability to understand.
- *Strong AI*:
  - Appropriately programmed computer is capable of conscious thought.
  - Machines can be made to think and have genuine understanding.

9.14 Turing test

- "Computing Machinery and Intelligence" by Alan Turing in 1950.
- Instead of asking:
  - "Can machines think?"
- Turing proposed a test:
  - "Can machines pass a behavioral test for intelligence?"
- He did not want to define intelligence and thinking.

Imitation game



Turing test

- Objections and responses (by Turing):
  - *"Head in the sand"* objection begs the question.
  - *Mathematical objection*: there are limits to what kind of questions a logic-based computer can answer.
  - *Mechanical objection:* encode all possible answers to a large memory of a fast computer.
  - Lady Lovelace objection: computers are incapable of originality $\rightarrow$ Turing: computers can still surprise us.
  - Informality objection: any rule-based system is predictable.
  - *Theological objection*: thinking is a property of immortal soul.

- Other criticism about the validity of the test:
  - Many humans may fail the test.
  - Simulation of conversational behavior is not an indication of intelligence.
  - Machine may be intelligent without being able to chat.
- Test of weak AI.
  - Does not say anything about consciousness or intentionality.
  - “Simulation is not the real thing.” (Gonick)

- Turing's prediction: by 2000, computer can carry on conversation with human for 5 minutes, and with 30% chance deceive the human.
- As of 2007, no computer has passed the test.
- Loebner prize (since 1990)
  - $2000 to the most humanlike chatterbot.
  - $25K to the program that judges cannot distinguish from human in a text-only test, and can convince the judges that the other one is computer.
  - $100K to the program that understands visual and auditory input in addition to text.

Will be given out once

- A reverse Turing test: CAPTCHA (2000)

  - Challenge-response test used in computing: user writes down letters of a distorted image.
  - Computer can generate and grade, but not solve.
  - Anyone entering the correct answer is assumed human.

9.15 Physical symbol system

- The system obeys the laws of physics.
- Not restricted to human symbol systems.
- Consists of
  - Set of entities, called symbols, that are physical patterns occurring as components in expressions (symbol structures).
  - A collection of processes that operate on the these expressions to produce other expressions: creation, modification, reproduction, deletion.
- Two central concepts:
  - *Designation* — The system can affect or be affected by an object through an expression that designates the object.
  - *Interpretation* — The process of executing a task designated by an expression.

- Physical symbol system (PSS) hypothesis (Newell & Simon, 1976):

  "A physical symbol system has the necessary and sufficient means for general intelligent action."

- General intelligence is

  "In any real situation, behavior appropriate to the ends of the system and adaptive to the demands of the environment ..."

- Result of successes of GPS.

- Physical medium – brain, computer, paper – does not matter.

- PSS is in core of strong AI.
- Qualitative specifies a general class of systems including those capable of intelligent behavior.
- Empirical — it can be tested if the class accounts for set of phenomena observed in real world.
- Appropriate for higher-level intelligence (e.g., playing chess) but not for commonplace intelligence.
- Frame problem originally demonstrated in this context.

9.16 Chinese room argument

- John R. Searle, 1980
- Not just a though experiment but the most (in)famous argument against strong AI.

- Searle's claim is that
  - Even if the answers given by the person in the room are indistinguishable from those of a native Chinese speaker, he does not understand a word of Chinese.
  - The room operates as a computer, performing computational operations on formally specified symbols.
  - Digital computer that operates on formal symbols cannot produce understanding, since there is no meaning involved.

9.17 Replies to the argument

- The systems reply
  - The understanding is ascribed to the whole room instead of to the individual in the room.
  - Searle's response: let's have the person memorize all the Chinese symbols and the rules for their manipulation.
- The robot reply
  - Let's put a computer inside a robot, and not just the symbol manipulation machinery, but also make it perceive and act; such a robot will exhibit genuine understanding and have mental states.
  - Searle's response: causal relations with outside world does not add anything to understanding if the cognition is based on symbol manipulation.

- The brain-simulation reply
  - Instead of designing a program that represents information, simulate the sequence of neural firings at the synapses of a Chinese person when understanding Chinese stories and answering questions.
  - Searle's response: This undermines the idea of strong AI; we don't need to know how the brain works in order to understand how mind works.
- The other-minds reply
  - How do you know that other people understand Chinese or anything? By their behavior.
  - Searle's response: the question is what is attributed to the system.

9.18 Searle's conclusions

- Could a machine think?
  - In principle no opposition to why machines could not understand Chinese or English.
  - Brains are a special kind of machine that can think.
  - All the machines with the same causal power as brain can think.
- Could something think or understand solely by virtue of being a computer with a right program?
  - No — formal symbol manipulation does not have intentionality.

- “Mind is to brain as program is to hardware.” Searle's counterarguments:
  - The same program can have several realizations.
  - Programs are formal (syntax), cognitive states are about the content (semantics).
  - Mental states are a product of the operation of brain, but programs are not a product of the computer.
- The computer simulation is not the real thing:
  - Fire or rainstorm
  - Love, pain, understanding