INF5390 – Kunstig intelligens

Foundations and Prospects

Roar Fjellheim

Outline

- The big questions
- Weak AI
- Strong AI
- Status of AI
- Prospects
- Summary

AIMA Chapter 26: Philosophical Foundations
AIMA Chapter 27: AI – The Present and Future

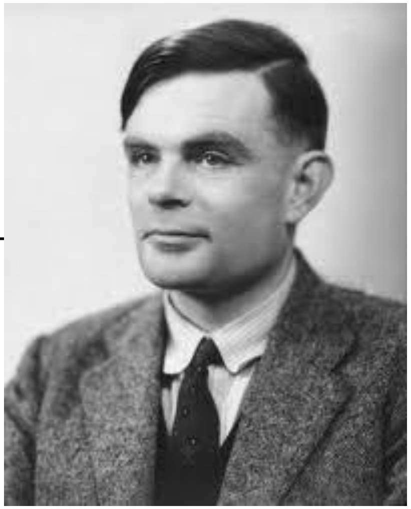
The Big Questions

- What does it mean to *think*?
- Are machines able to think?
- What is *intelligence*?
- Can machines be intelligent?
- What does it mean to be *conscious*?
- Can machines be conscious?
- What is *mind*?
- Can machines have mind?

Weak vs. strong AI

- ***Weak AI***
  - √ Machines can be made to act *as if* they are intelligent
- ***Strong AI***
  - √ Machines can be made that *are* intelligent, *have* minds, and *are* conscious

The Turing test



- In an attempt to answer the question “Can machines think?”, Alan Turing (1950) proposed the *Turing test* for intelligence
  - √ The computer shall have a conversation with an interrogator for 5 minutes and have a 30% chance of fooling the interrogator into believing it is human
- Turing believed that by year 2000, a computer with a storage of 10^9 units will pass the Turing test
  - √ So far, no computer has passed the test
- Such a machine will qualify as *weak AI* (“as if intelligent”)

Objections to intelligent machines

- Turing considered many objections to AI
  - Argument from *disability*
  - The *mathematical* objection
  - The argument from *informality*
- Disability: A machine can never do X
  - X = to be kind, friendly, make mistakes, have sense of humor, fall in love, do something really new, …
  - Counter: Many such “impossibility claims” are unsupported, and some can be refuted

Mathematical objections to AI

- An AI program is a *formal system* implemented on a computer, and subject to *theoretical limits*, e.g.
    - The *incompleteness theorem* (Gödel): In any formal system powerful enough to do arithmetic, there are true statements that cannot be proved
- Humans can *overcome* formal limits, e.g. by "meta-transfer" to other formalisms and are therefore inherently superior
- Counter-arguments
    - Computers are finite machines, and are strictly not subject to Gödel's theorem
    - Intelligent humans also suffer from inability to prove all true statements
    - The brain is a deterministic physical device (some argue against this) and subject to the same formal limits as as computer

Informality objection to AI

- Proposition (Dreyfus):
  - Human behavior is too complex to be captured by a simple set of *rules*
  - Since computers can only follow rules (can only do what the are told to), they cannot generate intelligent behavior on human level
- This critique is directed towards simple first-order logic rule-based systems without learning
  - "GOFAI - Good Old Fashioned AI"
- Modern AI includes other reasoning&learning methods
  - Generalization from examples
  - Supervised, unsupervised and reinforcement learning
  - Learning with very large feature sets
  - Directed sensing
- Thus, AI makes progress to overcome the critique

Strong AI - machine consciousness

- Even if machines can be made to act *as if* they are intelligent (weak AI), “real” machine intelligence must have *consciousness* (strong AI)
- The machine must be aware of its own *mental state* and actions, be aware of its own beliefs, desires and intentions
- Turing rejected this requirement, because we do not even know that other humans have consciousness, we can only observe their external behavior
- Many will nevertheless require strong AI before they accept a machine as intelligent

Can machines have mental states?

- *Functionalism* answer
  - If the computer provides same answer to a problem as a human would (same *function*), it must have the same internal mental state
- *Biological naturalism* answer
  - Mental states are high-level and *emergent* features that are caused by neural activity in the brain that cannot be replicated by other means

The mind-body problem

- Ancient question
  - How is *mind* (soul, consciousness) related to *body* (brain)?
- *Dualist* view
  - Mind and body are fundamentally different categories of existence
- *Materialist* view
  - “Brains cause minds” (Searle)
  - I.e. the brain is the “hardware” for the mind “software”
- Accepting the materialist view, can a machine have consciousness?

The Chinese room (Searle)

- Argument by Searle (1980)
  - Human ("CPU") with no knowledge of Chinese operates in a closed room with a rulebook ("program") and a stack of paper ("memory")
  - Human receives slips of paper with (for him non-intelligible) Chinese text, follows rules mechanically and returns sensible replies in Chinese
  - From the outside, it seems that the Chinese room behaves intelligently, yet the human has no idea of what he is responding to the inputs (just follows the rules)
- This demonstrates that a system that passes Turing test need not be intelligent or conscious

The Systems reply (McCarthy)

- The Chinese room argument relies on following claims
  - √ Certain kinds of objects are incapable of conscious understanding (in this case, Chinese)
  - √ The human, paper, and rule book are objects of this kind
  - √ If each of the objects is incapable of conscious understanding, then any system constructed from the objects is incapable of conscious understanding
  - √ Therefore there is no conscious understanding in the Chinese room
- In the “Systems reply” to Searle (McCarthy and others), the third claim is not accepted
  - √ If it was true, how could (conscious) humans be made of (unconscious) molecules?

Consciousness as emergent property

- In more recent work, Searle claims that consciousness is an *emergent property* of properly arranged neurons, and *only* (biological) neurons
- (Most) AI researchers agree that consciousness is an emergent property, but that the physical components underlying it can be neurons *or* electronic components *or* some other mechanism
- Searle’s argument is not more founded on “facts” than the opposite (AI) argument

Can the strong AI question be settled?

- Consciousness is not a well defined or well understood phenomena
- We do not know what kind of experiment can be used to determine consciousness in a computer
- Question could be settled if we discovered how consciousness can be *reduced* to other phenomena
- As no such reduction is known, the strong AI question will remain open

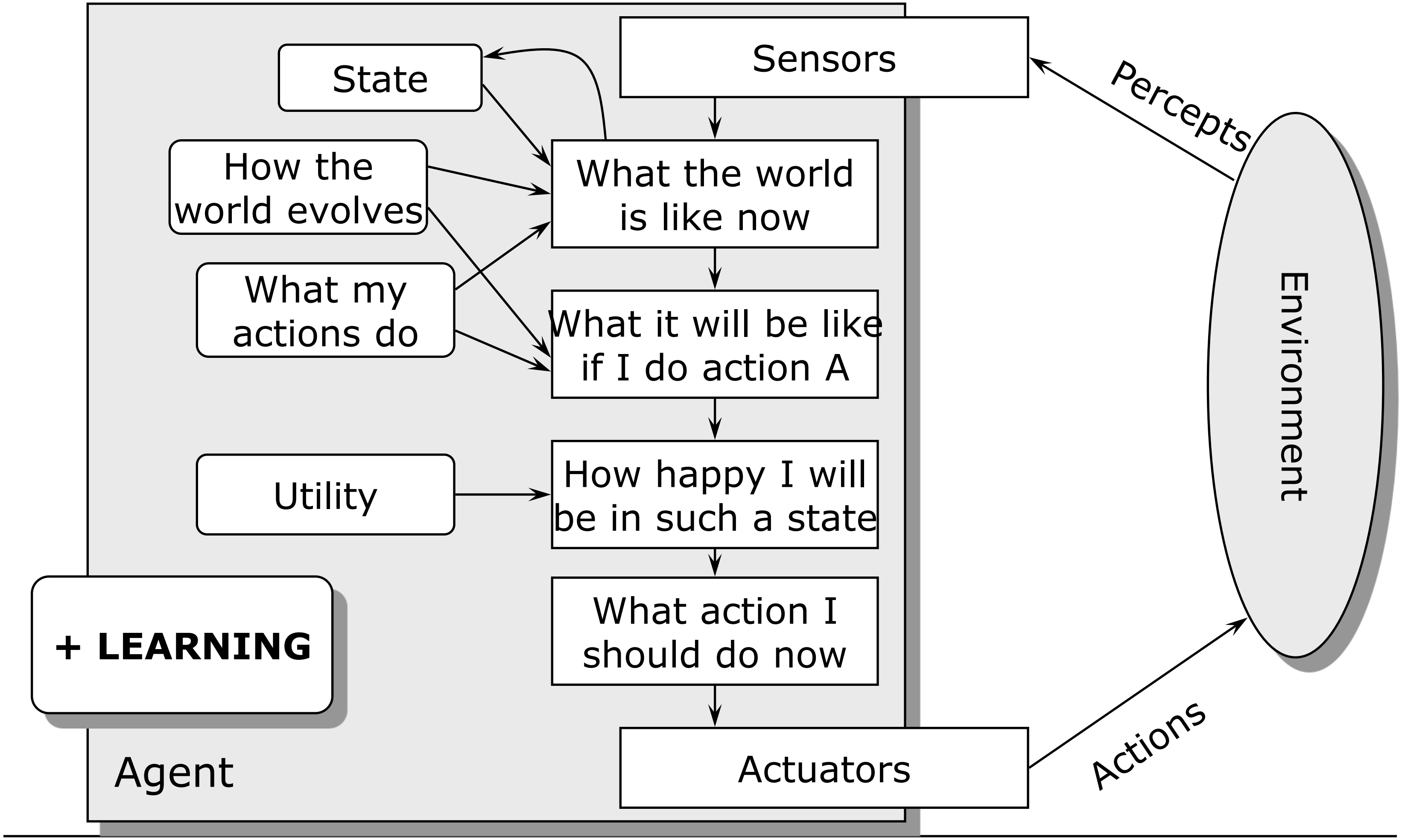
Tentative answers to some “big questions”

- Weak AI (machines can be made that act *as if* they are intelligent)
  - Many AI programs do in fact exhibit “intelligence”
  - Arguments against weak AI are needlessly pessimistic
- Strong AI (machines can be made that *are* intelligent and conscious)
  - Difficult to prove either impossibility or possibility of this claim
  - The answer is not important for further progress for (weak) AI

Recapitulation: AI as agent design

- The AI “project” can be seen as the design of *intelligent agents*
- Different agent designs are possible, from *reflex* agents to *deliberative* knowledge-based ones
- Different paradigms are being used: logical, probabilistic, “neural”
- Do we have the necessary tools to build *a complete, general-purpose agent*?

Model- and utility-based agent


State
Sensors
Percepts
How the world evolves
What the world is like now
What my actions do
What it will be like if I do action A
Environment
Utility
How happy I will be in such a state
+ LEARNING
What action I should do now
Agent
Actuators
Actions

State-of-the-art

- Interaction with the environment
  - Improved greatly in recent years: cameras, MEMS, ..
  - Dominant new environment: the Internet
- Keeping track of environment’s state
  - Perception and updating of internal representation
  - Filtering methods for tracking uncertain environments
  - Mostly low-level and propositional
  - Need to improve ability to recognize higher-level objects, relations, scenes, etc.

State-of-the-art (cont.)

- Evaluate and select actions
  - √ Simple methods for planning and deciding exist
  - √ Real-world complexity require strong abstraction ability (hierarchies)
  - √ Great deal of development is needed
- Utility as expression of preference
  - √ MEU is sound in principle, but depends on realistic utility functions
  - √ Need to extract utility information from humans to guide agents

State-of-the-art (cont.)

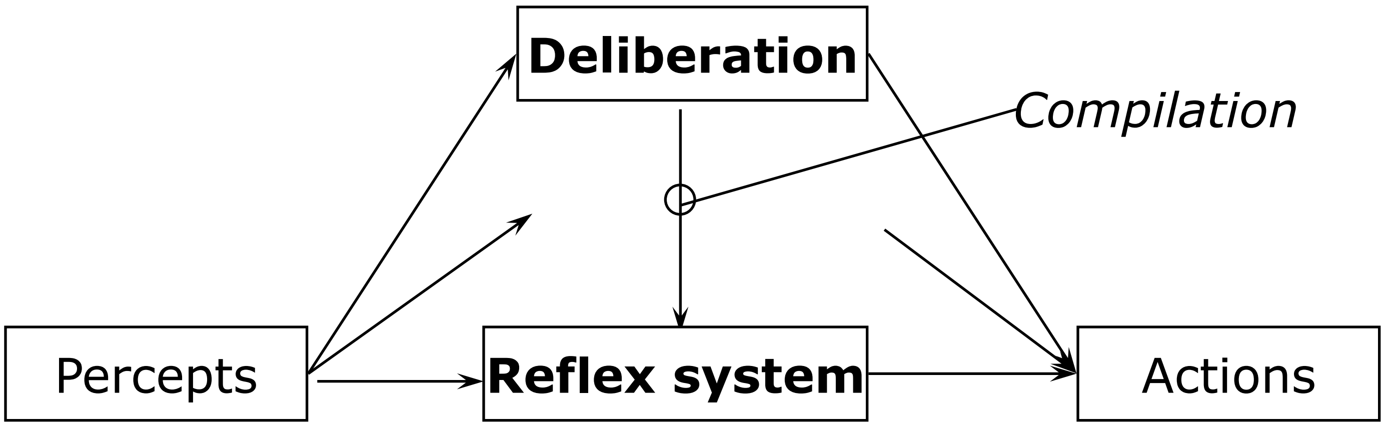
- Learning capabilities
  - Basic learning technology has progressed rapidly in recent years, sometimes with abilities that exceed human learning ability
- However, little progress on how to learn *higher level* concepts from lower level (input) concepts
  - Without such generalization ability, agents must be taught manually by humans

Uneven status of AI disciplines

- Some parts of AI are *mature*, and agents can be built that outperform humans in these areas
  - E.g.: Game playing, logical inference, theorem proving, planning, diagnosis
- Other parts of AI are *evolving*, where progress is being made
  - E.g.: Learning, vision, robotics, natural language understanding

Hybrid agent architecture

- Ability to incorporate different types of reasoning and decision making (from reflex to deliberation)
- Learning from experience (compiling)



Control of agent deliberation

- Real-time AI
  - Agents in the real world must act in real-time
- Anytime algorithms
  - Have an answer ready at all times, improve if more time available
- Decision-theoretic metareasoning
  - Use value of information to reason about which computation to perform
- Reflective architecture
  - Apply same kind of reasoning to internal decision-making as to external decision-making

AI as rational agents – right direction?

- Perfect rationality
  - √ Agent always does the right thing
  - √ Not feasible in non-trivial domains
- Calculative rationality
  - √ Will *eventually* do the right ting, but must be "short-circuited"
  - √ Underlies much of current AI
- Bounded rationality
  - √ Theory for how "real" agents solve problems
  - √ Satisficing: Deliberate only until answer is "good enough"
- Bounded optimality
  - √ Agent does best possible given its computational resources
  - √ Offers best promise for *strong theoretical foundation for AI*

If AI succeeds ...

- Intelligent agents, autonomous or working on behalf of humans: Who is responsible?
- AI impact on work and leisure, quality of life: Will it be positive or negative?
- AI impact on politics and power, governments and citizens: Who will gain and who will lose?
- If machines with high level intelligence develops, will they have rights? Relationship to humans?
- Will machines eventually supersede humans …?