Philosophical Foundations

- Weak AI
  - claim: computers can be programmed to act *as if* they were intelligent (*as if* they were thinking)
- Strong AI
  - claim: computers can be programmed to think (i.e., they *really are* thinking)
- Most AI researchers assume Weak AI is true and Strong AI is irrelevant

Is Weak AI possible?

- Traditionally posed as the question “Can Machines Think?”
  - problem: This depends on customary use of the word “think”
  - compare:
    - Can machines fly?
    - Can machines swim?

Flying and Swimming

- In English, machines can fly but they do not swim
- In Russian, machines can do both

Arguments against Weak AI

- Turing's famous paper "Computing Machinery and Intelligence" (1950)
  - Argument from disability
  - Argument from mathematics
  - Argument from informality

Argument from Disability

- A machine will never be able to…
  - 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.
- But nowadays, computers do many things that used to be exclusively human:
  - play chess, checkers, and other games, inspect parts on assembly lines, check the spelling of documents, steer cars and helicopters, diagnose diseases, and 100's of other things. Computers have made small but significant discoveries in astronomy, mathematics, chemistry, minerology, biology, computer science, and other fields

What about tasks involving "judgement"?

- Meehl (1955): linear regression is more accurate than trained experts at predicting student success in training programs or criminal recidivism
- The essay questions on the GMAT are graded by computer and agrees with human graders 97% of time (which is about the same as human graders agree with each other)
- But there are many difficult tasks still to be conquered…

The Mathematical Objection

- Godel's theorem shows that computers are mentally inferior to humans because machines are formal systems (limited by the theorem) but humans have no such limitation (J. R. Lucas, 1961)

Godel's Theorem

- Let F be a formal axiomatic system that is powerful enough to do arithmetic
- Then it is possible to construct a "Godel sentence" G(F) with the following properties
  - G(F) can be written as a sentence in F
  - G(F) cannot be proved using the axioms in F
  - but if F is consistent, then G(F) is true
- "There are true sentences that cannot be proved"

Replies to the Mathematical Argument

- Godel's theorem applies to Turing machines, but real computers are NOT Turing machines, so the theorem does not apply to them
- There are similar sentences that apply to people
  - "J. R. Lucas cannot consistently assert that this sentence is true."
    - If he asserts it, then he is contradicting it.
    - But this doesn't lead us to believe he is unintelligent.
- How do we know that people aren't subject to Godel-theorem type limitations? What evidence do we have?
  - People have many limitations; people are inconsistent
  - A proof that people are not subject to Godel's theorem would itself contain a formalization of this unformalizable capability, and therefore contradict itself.

Informality

- Claim: Human behavior is far too complex to be captured by any simple set of rules
- Computers can do no more than follow rules
- therefore: they cannot generate behavior as intelligent as humans
  - Dreyfus (1972): "What Computers Can't Do"
  - Dreyfus (1992): "What Computers Still Can't Do"
  - Dreyfus & Dreyfus (1986): "Mind Over Machine"

The Qualification Problem

- It is impossible in formal logic to write down all of the conditions that must be true in order for a certain action to succeed
  - picking up a book won't succeed if
    - the book is glued to the table
    - the book is slippery, your hand is covered with butter
    - a meteor destroys the book just before you touch it
    - an earthquake hits and causes you to fall over
    - etc.

Probability to the Rescue (partially)

- Our lack of knowledge about all of these possible strange cases can be summarized in a probabilistic statement
  - The probability that you will succeed in picking up the book is 0.999
  - We can reason soundly with this statement
    - (whereas we could *not* reason soundly in logic)
- But still we need some form of "variable resolution" model that allows us to consider these strange possibilities as appropriate
  - Unsolved problem, but perhaps not impossible(?)

Dreyfus has made a career out of being an AI Skeptic

- Russell says Dreyfus' book should have been entitled
  - "What first-order logic rule systems without learning can't do"
- With each new book, Dreyfus has shifted more from being a critic to being an AI researcher himself

Strong AI: Can Machines Really Think?

- Even if a machine passes the Turing test, it will just be a simulation of thinking, not *real* thinking

Simulations and Realities

- Cases where Artificial = Real
  - Synthetic Urea was really urea (Wohler, 1848)
  - Artificial sweeteners are really sweeteners
  - Artificial Insemination (the “other AI”) is really insemination
- Cases where Artificial $\neq$ Real
  - Artificial flowers are not flowers
  - Artificial Chateau Latour win is not real Chateau Latour (even if chemically identical)
  - Artificial Picasso painting is not a real Picasso painting no matter how wonderful
- Is AI like Urea or like Artificial Picasso?

Computer Simulations

- computer simulation of addition is addition
- computer simulation of chess is chess
- Is computer “simulation” of reasoning reasoning?

Functionalism

- Claim: There is a level of abstraction below which the specific implementation doesn't matter
  - A computer and a brain could both implement this same level of abstraction, and therefore be "isomorphic"
- Strong AI claims that this level exists
- Therefore, if human brains are "thinking", then so are properly-programmed computers

Two Thought Experiments

- Brain Prosthesis Experiment
- Chinese Room

Brain Prosthesis

- Suppose we can develop artificial digital neurons that can perfectly mimic the behavior of each neuron in the brain
- Suppose we replace your neurons one-by-one with these artificial neurons
- Claim: You will notice no difference in your conscious experiments
- Therefore: The functional level of abstraction is at-or-above the neuron level

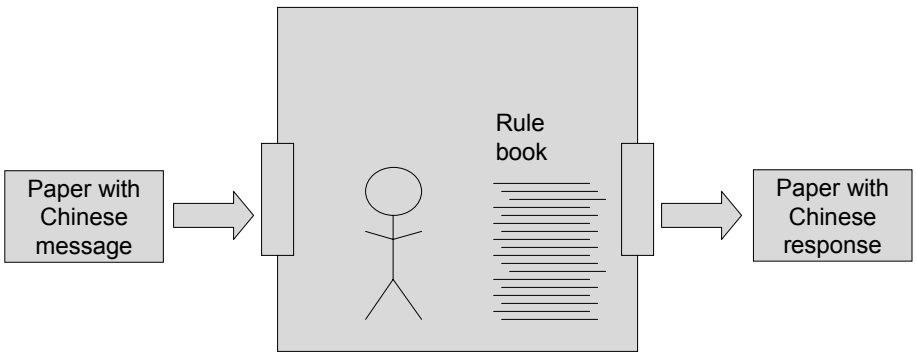
Searle (philosopher of mind) believes that consciousness will be lost

- "You find to your total amazement, that you are indeed losing control of your external behavior. You find, for example, that when doctors test your vision, you hear them say "We are holding up a red object in front of you, please tell us what you see." You want to cry out, "I can't see anything, I'm going totally blind." But you hear your voice saying in a way that is completely out of your control, "I see a red object in front of me"…Your conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same"

Russell says

- The resulting artificial brain will still produce all observable phenomena of consciousness
  - e.g., if we ask it "How do you feel?" it will answer "I feel fine. I must say I'm a bit surprised, because I believed Searle's argument."
- Any explanation of these observable manifestations of consciousness would apply equally well to the biological brain
- Therefore, either
  - The causal mechanisms of consciousness have been replicated, or
  - The manifestations of consciousness are "epiphenominal" – they have no causal connection to true consciousness

The Chinese Room (Searle)



The person applies the rules to convert the input paper into the output paper. But the person does not "understand" the rules nor does he "understand" Chinese

Chinese Room (2)

- Searle's argument
  - The person inside the room does not understand Chinese
  - The rule book, being just many sheets of paper, does not understand Chinese
  - Therefore, there is no understanding of Chinese going on
- Running "the right program" does not produce understanding

The Systems Reply

- Intelligence is an emergent property of the entire system
  - If you ask the CPU whether it can take cube roots, the answer is no.
  - The program, of course, can't take cube roots unless it is executed by the CPU
  - But the computer as a whole can take cube roots
- The computer executing the right program could create isomorphic states
- Searle does not have a good response to this

Bounded Optimality Undermines Functionalism

- In this class, we have defined intelligence as "exhibiting the best performance attainable on the given hardware"
  - This implies that the implementation *does matter* unless it exhibits the same resource tradeoffs (space, time, energy, etc.), which is highly unlikely!
  - Therefore, artificial (silicon-based) intelligence and human (neuron-based) intelligence are likely to make different tradeoffs and exhibit different strengths and weaknesses
  - They will not share "isomorphic" states.
  - We will have to leave it to future humans to decide whether to call their internal processing "thinking".

Will Human-Level AI Ever Exist?

- Do digital circuits exhibit the right tradeoffs? Not clear yet
- Is there a market for HLAI?
- I think there will be much more of a market for superhuman intelligence in niche applications, just as there is now