AI and Philosophy

Gilbert Harman

Tuesday, December 4, 2007

My Background

- Web site `http://www.princeton.edu/~harman`
- Philosophy
- Early Work in Computational Linguistics (including MT Lab at MIT)
- Cognitive Science
- PSY 237/PHI 237 "The Psychology and Philosophy of Rationality," taught with Eldar Shafir and Philip Johnson-Laird.
- PHI 218/ELE 218 "Epistemology and Learning Theory," taught with Sanjeev Kulkarni.
  - Our new book, *Reliable Reasoning: Induction and Statistical Learning Theory* (MIT Press, 2007).

A Philosophical Question about Personal Identity

- What is it to be a person? What is the difference between people and other animals?
- Classical view: "Man is the rational animal."
  - Human beings are animals that think, intelligent animals.
  - To understand what it is to be a person is to understand what it is to be a rational being.
- But maybe there are rational beings that are not animals
  - Gods, angels
  - ETs
  - Robots, computers, ...

Understanding through Analogies

- Atomism.
- Wave theory of sound.
- Current theory of electricity.
- Particle and wave theories of light.
- Models of mind
  - Mechanical toys of the 16th Century suggested a person might be a machine.
  - Developments in AI suggest models of human intelligence.

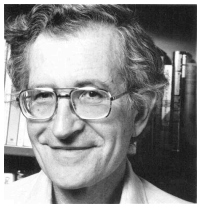
Rene Descartes 1596-1650

Descartes' Dualism

- Background physics: contact mechanics, billiard ball models.
  - No action at a distance
  - No fields: gravity, magnetism, electricity.
- Argument that mind is not explainable mechanically.
  - Mind involves thought, association of ideas, and reasoning.
  - Other animals can perhaps be explained mechanically. But not people. People are basically their minds, attached to their bodies.
  - Explanation of mind is different from explanation of bodies. No mere mechanical explanation of mind.
  - Implication?: Perhaps minds can survive the destruction of the bodies.

Conversation

- In particular, Descartes argued that it is inconceivable that mechanical principles could explain ordinary human conversation, which has features of *novelty* and *appropriateness*.
- Compare Turing Test in Alan Turing, "Computing Machinery and Intelligence".
- According to the linguist Noam Chomsky, explaining ordinary conversation is a *mystery* rather than a *problem*.
- Human beings are the language using animals.



Animals

- According to Descartes: animals are in principle explicable mechanically.
- They do not think or reason in the way people do. They do not act on reasons.
- They do not have immortal souls.
- They do not have language.
- Some followers of Descartes went around kicking dogs in order to show their allegiance to dualism.

Interaction

- Mind and body must interact in perception and in action.
- This raises the problem: how? How can something that is not a body have an effect on body? And vice versa?
- Descartes argues that the point of interaction between the two realms occurs in a certain gland in the brain, the pineal gland.
- But that does not really address the problem.

Developments in Physics

- Later developments: changes in physics allowed for nonmechanical effects and action at a distance.
- This opens up new possibilities for mind body interaction.
- Maybe the effect of mind on body is like the effect of gravity or the effect of magnetism.
- Perhaps a mind is something like a field.
- Quantum physics suggests additional possibilities.
- ESP?

20th Century Analogies for Mind

- Telephone switchboard analogy.
- Information theory. Mind as an information processing system.
- Flow charts in programming: Psychological theories as flow charts.
- Logic programming: thinking as theorem proving.
- Computer: mind as a computer; person as a computer in a robot.
- Subroutines and modularity, psychological modularity: perceptual systems, language systems, motor systems, face-recognition system.
- Expert systems: intelligence as expertise.
- Pattern recognition and computational learning theory: psychological learning as pattern recognition.

Rethinking Descartes' Argument against a Mechanical Explanation of Mind

- Suppose the mind is the brain, which is like a computer in a robot body.
- One possibility is that the brain functions in terms of principles that go beyond Descartes' mechanics.
- To be sure, in principle it is possible to have computers that operate completely mechanically (Babbage).
- But there may be a difficulty with supposing the brain is a mechanical computer, having to do with considerations of size and speed.

Anti-Dualism: Three Sorts of Theory

- Behaviorism
- Double-Aspect Theory
- Functionalism

Behaviorism

- Behaviorism equates mental states and occurrences with behavior.
  - Including behavioral tendencies and dispositions.
  - Being magnetic is an example of a disposition.
  - Being magnetizable is a second-order disposition.
- Turing on computational intelligence
- Problem about stoicism.

Double Aspect Mind-body Identity Theory

- Mind-body identity theory.
  - Pain is activity in certain C-fibers.
  - Compare
    - Lightning is an electrical discharge.
    - Water is H_2O
- Double-Aspect version: Some physical events (from the outside) are mental events (from the inside).

"Functionalism"

- Things can be identified in terms of function.
- It does not matter what they are made of.
  - Artifacts: A pencil functions as a writing instrument of a certain sort.
  - Organs: A heart functions to pump blood.
- Mental events can be identified with whatever physical events have the relevant functional (causal) properties.

Mind-Body Problem (Inside/Outside)

- I know what it is like to be me from the inside.
- I want to know how what it is like to be me from the inside fits with what it is like to be me from the outside as revealed to other people, or science
- Related issue: knowing what it's like to be you via an understanding of you from the outside.
- The problem of other minds: how do I know that there is something it is like to be you?

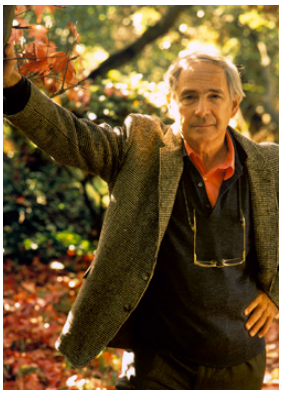
Two Kinds of Understanding?

- Claim: the mind-body problem is an illusion that arises through ignoring two kinds of understanding
  - The method of the sciences: understanding things from the "outside" by seeing them as instances of general laws.
  - The method of the humanities: understanding cognitive and social phenomena from the "inside" by relating them to your own experiences—by translating them into your own terms.
- The one sort of understanding is not enough for the other sort.

Aren't You Special?

- People are essentially unique beings in some hard to articulate way.
  - Doesn't it seem essential to you that you are different from all other people in an extremely important respect?
- If you case to function, you die, you go out of existence.
  - Computers and robots are different.

John Searle



John Searle's Chinese Room Argument against Functionalism

- A system might behave as a speaker of Chinese and contain events with the right functional properties without understanding Chinese.
- A computer simulating a Chinese speaker does not understand Chinese.
- Computer representations do not have *intrinsic* content ("intentionality")
  - A person's thoughts have intrinsic intentionality.
  - The writing on this slide has derived intentionality.
  - The representations of my bank balance in my computer program have only derived intentionality.
  - So do the representations in the Chinese understanding program.

Responses to Searle's Argument

- Searle confuses the CPU with the whole system: the whole system understands Chinese even if the CPU does not.
- The distinction between derived and intrinsic intentionality is suspect. Maybe a person's thoughts have only derived intentionality.

Discussion ...