Philosophy 101

Searle's Chinese Room Argument

March 6, 2014

Review

Strong & Weak AI

Strong & Weak AI

Strong AI: thinking is constituted by the manipulation of formal symbols, such as occurs in a computer program

Weak AI: thinking may be modeled by formal symbol systems, such as computer programs

The Turing Test

The Turing Test

1. For some arbitrary time period, there will be no discernable difference between the linguisitc behavior of a person and that of a machine
2. If there is no discernable difference in linguistic behaviour between man and machine, then there is no reason to think that there is any underlying difference in the causes of that behaviour
3. ∴ If we are willing to say that it is intelligent thought that is the cause of the linguistic behavior in the person we should be willing to say the same thing about the machine

Strong AI & the Turing Test

- Any computer that can pass the Turing Test for arbitrarily long periods of time will, according to strong AI, qualify as a thinking machine

Syntax & Semantics

Syntax & Semantics

Syntax: the formal or structural features of a symbol system which determine which expressions are legitimate or well-formed members of the system and which are not

- The syntax of English (its grammar) requires that all complete sentences have a noun phrase and a verb phrase
  - 'John goes to school' vs. 'school John to goes'

Syntax & Semantics

Semantics: The system of meanings assigned to a symbol system, given by determining the referents of the symbols and the truth conditions of symbol strings

- 'Schnee' refers to snow
- 'weiß' refers to the property of being white
- 'Schnee ist weiß' is true just in case snow is white

The Chinese Room Argument

The Chinese Room Argument

1. Programs are purely formal (syntactic)
2. Human minds have mental contents (semantics)
3. Syntax by itself is neither constitutive of, nor sufficient for, semantic content
4. ∴ Programs by themselves are not constitutive of nor sufficient for minds

What Does the Argument Intend to Prove?

What Does the Argument Intend to Prove?

- You can't get semantic content from syntax alone
- A system must have more than purely syntactic properties in order to possess intentional states

Objections

Two Objections

Two Objections

1. The 'Systems' Objection
2. The 'Implementation' Objection

The 'Systems' Objection

The 'Systems' Objection

- Perhaps the person *in* the Chinese room does not understand Chinese but the *Chinese Room itself* understands Chinese
- Since the Chinese room is the proper analogue to the computer program, and not the person *in* the Chinese room, Searle's example proves nothing

The 'Systems' Objection

Searle's Reply

My response to the systems theory is quite simple: let the individual internalize all of these elements of the system. He memorizes the rules in the ledger and the data banks of Chinese symbols, and he does all the calculations in his head. The individual then incorporates the entire system. There isn't anything at all to the system that he does not encompass. We can even get rid of the room and suppose he works outdoors. All the same, he understands nothing of the Chinese, and a fortiori neither does the system, because there isn't anything in the system that isn't in him. If he doesn't understand, then there is no way that the system could understand because the system is just a part of him.

The 'Systems' Objection

Pryor's Rebuttal (I) Searle's argument is invalid

> *Searle: "he understands nothing of the Chinese, and [therefore] neither does the system, because there isn't anything in the system that isn't in him"*

The 'Systems' Objection

Pryor's Rebuttal (I) Searle's argument is invalid

> *Searle: "he understands nothing of the Chinese, and [therefore] neither does the system, because there isn't anything in the system that isn't in him"*

- This is a bad inference—compare:

The 'Systems' Objection

Pryor's Rebuttal (I) Searle's argument is invalid

Searle: "he understands nothing of the Chinese, and [therefore] neither does the system, because there isn't anything in the system that isn't in him"

- This is a bad inference—compare:

Searle doesn't weigh 3 pounds, and therefore neither does his heart, because there is nothing in his heart that isn't in him

The 'Systems' Objection

Pryor's Rebuttal (I) Searle's argument is invalid

- the form of inference Searle uses here doesn't generalize to other inferences with the same kind of form
  - leaves open the possibility that the particular argument Searle makes here is true

The 'Systems' Objection

Pryor's Rebuttal (II) Internalization is irrelevant

> *Searle: "If he doesn't understand, then there is no way that the system could understand because the system is just a part of him."*

The 'Systems' Objection

Pryor's Rebuttal (II) Internalization is irrelevant

- Consider a software emulator
  - allows one operating system to run 'on top of' another using the same hardware
    - Mac computers can emulate the Windows OS

The 'Systems' Objection

Pryor's Rebuttal (II) Internalization is irrelevant

- Assume a Mac runnning its OS *plus* an emulation of Windows OS
  1. The Windows OS is integrated or incorporated into the Mac OS
  2. Nevertheless, the states of the 'incorporated' Windows OS are in many ways independent of the Mac OS and its states

  - Windows may crash and become unresponsive, while the Mac software (including the emulator) keeps running
  - Windows might be treating Internet Explorer as the frontmost, active program; but–if you don't have the emulator software active in your Mac–the Mac software could be treating Safari as its frontmost, active program

The 'Systems' Objection

Pryor's Rebuttal (II) Internalization is irrelevant

> *when Jack memorizes all the instructions in the Chinese book, he becomes like the Mac software, and the Chinese room software becomes like the emulated Windows software. Jack fully incorporates the Chinese room software. That does not mean that Jack shares all the states of the Chinese room software, nor that it shares all of his states. If the Chinese room software crashes, Jack may keep going fine. If the Chinese room software is in a state of believing that China was at its cultural peak during the Han dynasty, that does not mean that Jack is also in that state. And so on. In particular, for the Chinese room software to understand some Chinese symbol, it is not required that Jack also understand that symbol.*

The 'Systems' Objection

Pryor's Rebuttal (II) Internalization is irrelevant

- Problem 2: 'Internalizing' the Chinese room program is irrelevant
  - two programs running on the same hardware need not share all of the same (or any of the same) states

The 'Systems' Objection

Summary of Pryor's Rebuttals:

The 'Systems' Objection

Summary of Pryor's Rebuttals:

1. Searle's argument is invalid
   - the form of inference Searle uses here doesn't generalize to other inferences with the same kind of form in a way that preserves truth
2. 'Internalization' is irrelevant
   - two programs running on the same hardware need not share all of the same (or any of the same) states

The Implementation Objection

The Implementation Objection

- We need to distinguish between a *program* and an *implementation of the program*

The Implementation Objection

Programs are abstract computational objects and are purely syntactic. Certainly, no mere program is a candidate for possession of a mind. Implementations of programs, on the other hand, are concrete systems with causal dynamics, and are not purely syntactic. An implementation has causal heft in the real world, and it is in virtue of this causal heft that consciousness and intentionality arise. It is the program that is syntactic; it is the implementation that has semantic content. (Chalmers 1996, 327)

The Implementation Objection

Chalmers's Parody:

The Implementation Objection

Chalmers's Parody:

1. Recipes are syntactic.
2. Syntax is not sufficient for crumbliness.
3. Cakes are crumbly.
4. ∴ Implementing a recipe is insufficient for a cake.

The Implementation Objection

A recipe implicitly specifies a class of physical systems that qualify as implementations of the recipe, and it is these systems that have such features as crumbliness. Similarly, a program implicitly specifies a class of physical systems that qualify as implementations of the program, and it is these systems that give rise to such features as minds. (Chalmers, 327)