Philosophy 101

Representation, AI, and the Mind

May 27, 2014

Philosophy 101

Representation, AI, and the Mind

Review

Representation

Can Machines Think?

Searle & the Chinese Room

Objections

Review

Quiz

1. What is a statement?
2. T/F: The modal argument assumes that phenomenal zombies are possible
3. T/F: Physicalism denies that phenomenal zombies are possible
4. T/F: Lewis defends the Hypothesis of Phenomenal Information
5. Name one of the abilities discusses by Lewis's "Ability Hypothesis"

Representation

What is a Representation?

Representation: something that represents something (either itself or something else)

- goes proxy; stands for; symbolizes something
- refers to something; is accurate/inaccurate; is true/false

Examples of Representational Kinds

- Pictorial Representation
- Linguistic Representation
- Mental Representation

Two Questions

1. How does a representation represent?
2. Are some kinds of representation more fundamental than others?

Test Cases

1. How does a representation represent?
    - resemblance
2. Are some kinds of representation more fundamental than others?
    - pictorial representation

Test Cases

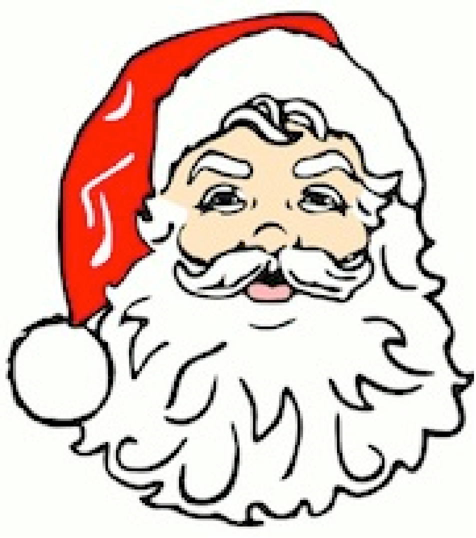


Figure: Who do I resemble?

Test Cases – Resemblance

- Resemblance is neither necessary nor sufficient for representation
  - not necessary: pictures can *represent* non-existent objects but they cannot *resemble* non-existent objects
  - not sufficient: everything resembles something but not everything represents something

Test Cases – Pictures

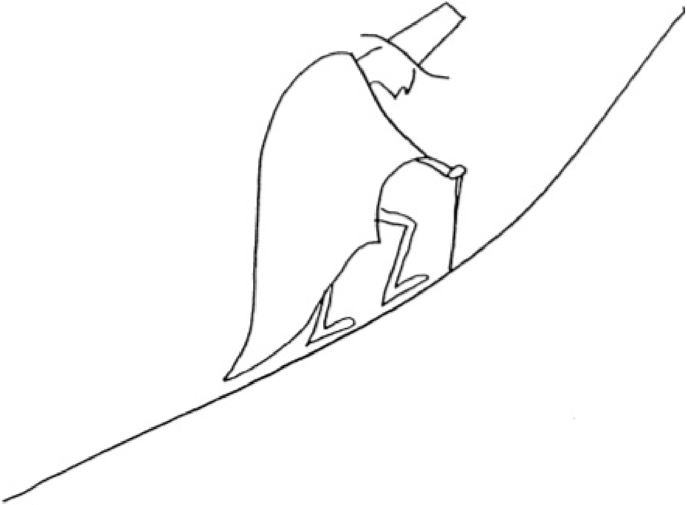


Figure: Walking Uphill or Sliding Down?

Test Cases – Pictures

1. Interpretation
2. Logical relations
   - if…then…
   - …and…
   - either…or…
   - not…

Test Cases – Pictures

- Pictorial representations:
    - require interpretation
    - cannot represent logical relations

Test Cases

- Linguistic Representation
    - convention
    - expression of ideas

Test Cases

words, like pictures, do not represent in themselves ('intrinsically'). They need interpreting – they need an interpretation assigned to them in some way. But how can we explain this? The natural answer, I think, is that interpretation is something which the mind bestows upon words. (Crane, p. 22)

Fundamentality

1. Any representational system that depends on interpretation or convention is not fundamental
2. Linguistic and pictoral representation depend on interpretation and convention
3. ∴ Linguistic and pictoral representation are not fundamental

Is Mental Representation Fundamental?

Mental representation is not conventional

1. Conventions rely on the intentions of a subject or subjects
2. Intentions are a kind of mental representation
3. ∴ Convention cannot be used to *explain* mental representation

Is Mental Representation Fundamental?

Mental representation does not depend on interpretation

1. If mental representation required interpretation then we would need to be able to think about the interpretations in some way, in order to use them as interpretations
2. But thoughts are a form of mental representation, so we would need a further interpretation to make sense of those thoughts
3. But a further thought would be needed to think *that* interpretation, etc.
4. ∴ we would have an infinite regress of thoughts and interpretations of them
5. ∴ Mental representations do not require interpretation

Can Machines Think?

Two Questions

1. Can a physical system capable of performing certain functions think?
2. Can a sufficiently sophisticated computer program think?

Two Questions

2. Can a sufficiently sophisticated computer program think?
   - Is the mind to the brain like software is to hardware?

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 Imitation Game

- Can you guess, using a series of questions, which of two conversation partners is a machine and which a human?
- Questions may be of all kinds:
    - what's your name
    - what's your favorite color?
    - what does the smell of freshly cut grass remind you of?

The Turing Test

I believe that in about fifty years' time it will be possible to programme computers…to make them play the imitation game so well that an average interrogator will not have more than 70 percent chance of making the right identification after five minutes of questioning…I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted. (Alan Turing)

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

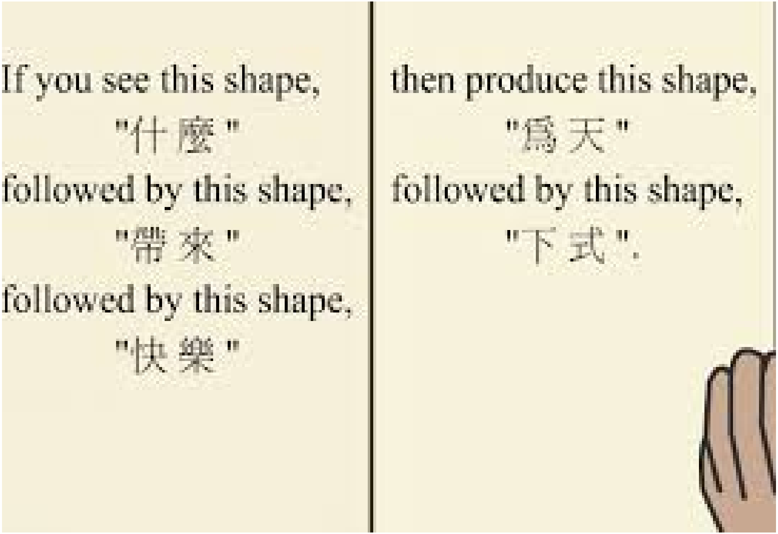
- Any computer that can pass the Turing Test for arbitrarily long periods of time will, according to strong AI, qualify as a thinking machine

Searle & the Chinese Room

The Chinese Room Argument

suppose I am placed in a room containlng baskets full of Chinese symbols. Suppose also that I am given a rule book in English for matching Chinese symbols with other Chinese symbols. The rules identify the symbols entirely by their shapes and do not require that I understand any of them. The rules might say such things as, "Take a squiggle-squiggle sign from basket number one and put it next to a squoggle-squoggle sign from basket number two." Imagine that people outside the room who understand Chinese hand in small bunches of symbols and that in response I manipulate the symbols according to the rule book and hand back more small bunches of symbols.

The Chinese Room Argument



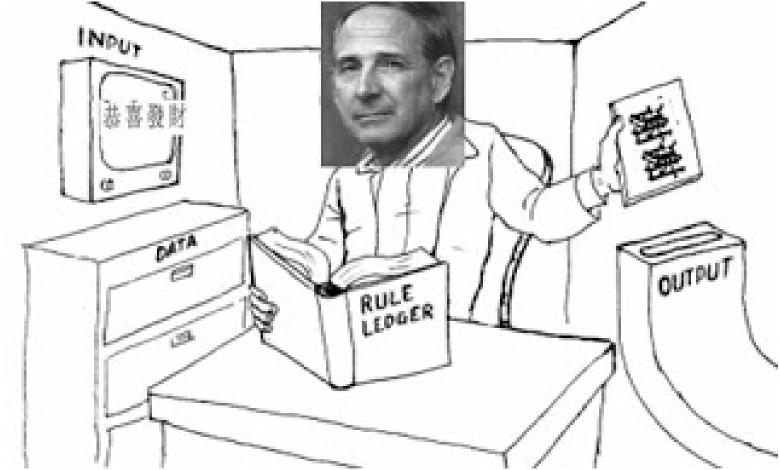
If you see this shape,
"什麼"
followed by this shape,
"帶來"
followed by this shape,
"快樂"

then produce this shape,
"爲天"
followed by this shape,
"下式".

The Chinese Room Argument

Now, the rule book is the "computer program." The people who wrote it are "programmers," and I am the "computer." The baskets full of symbols are the "data base," the small bunches that are handed in to me are "questions" and the bunches I then hand out are "answers."

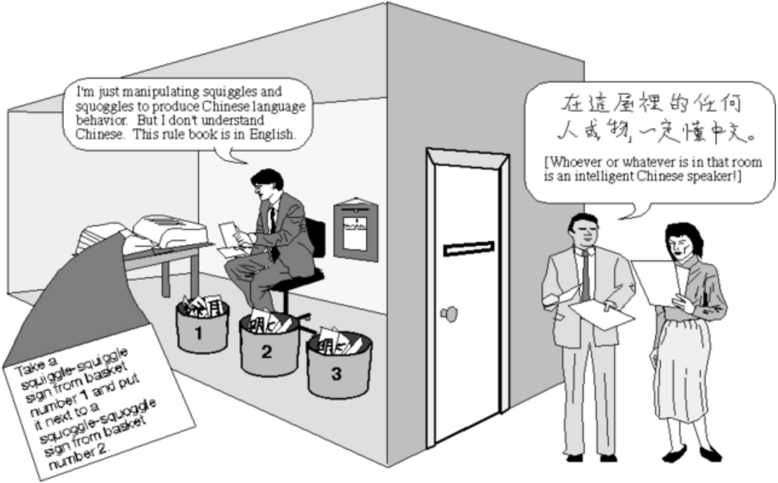
The Chinese Room Argument



The Chinese Room Argument

Now suppose that the rule book is written in such a way that my "answers" to the "questions" are indistinguishable from those of a native Chinese speaker. For example, the people outside might hand me some symbols that unknown to me mean, "What's your favorite color?" and I might after going through the rules give back symbols that, also unknown to me, mean, "My favorite is blue, but I also like green a lot." I satisfy the Turing test for understanding Chinese. All the same, I am totally ignorant of Chinese. And there is no way I could come to understand Chinese in the system as described, since there is no way that I can learn the meanings of any of the symbols. Like a computer, I manipulate symbols, but I attach no meaning to the symbols. (Searle, 26)

The Chinese Room Argument


I'm just manipulating squiggles and squoggles to produce Chinese language behavior. But I don't understand Chinese. This rule book is in English.
在這屋裡的任何人或物，一定懂中文。
[Whoever or whatever is in that room is an intelligent Chinese speaker!]
1
2
3
Take a squiggle-squiggle sign from basket number 1 and put it next to a squoggle-squoggle sign from basket number 2.

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

Syntax & Semantics

Syntax: the formal or structural features of a symbol system which determine which expressions are legitimate 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

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

1. The ‘Systems’ Objection
2. The ‘Implementation’ 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"

- 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:

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

- 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:

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)