

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



علوم شناختی

جلسه ۱۰ (ب)

# مباحثه‌ی اتاق روسی

## The Russian Room Argument

کاظم فولادی قلعه  
دانشکده مهندسی، دانشکدگان فارابی  
دانشگاه تهران

<http://courses.fouladi.ir/cogsci>

# PART 2: MODELS AND TOOLS



# Chapter 4: Physical Symbol Systems and the Language of Thought



# Chapter 4.3: The Russian room argument



# Lines of attack

---

- **Frame problem**  
Presents difficulties for the idea of representing knowledge symbolically
- **Russian room argument**  
Challenges the syntactic assumption at the heart of the PSSH

# Frame problem

---

## Original version (McCarthy and Hayes 1969)

How can a formal system represent the changes brought about by an action without explicitly representing all the things that the action does not bring about?

# Broader versions of frame problem

---

Some theorists have argued that the frame problem poses an in principle objection to the PSSH

- (Alleged) impossibility of formalizing commonsense reasoning
- Often accompanied by emphasis on “situatedness” and “embodiment” of real cognitive agents

# Assessment?

---

- It is hard to know how to assess these arguments without explicit impossibility proofs
- The real test comes with the alternative models proposed
  - Connectionist models of knowledge representation
  - Embodied/situated AI



# Syntax

---

- Physical symbol structures are purely syntactic
  - The symbols do not have any intrinsic meaning
  - Nor do the expressions built up out of them
  - The operations on physical symbols are sensitive only to the “shape” of those symbols
    - Formal rules, like the rules of a logical calculus

# From syntax to semantics

---

- One can specify a complete machine table for a TM without saying anything about what it is intended to represent (its *intended interpretation*)
- The machine table just specifies what the appropriate transitions are for any possible combination of inputs and states
- But if we assign meanings to the symbols then we can interpret the machine as carrying out specific calculations

# A sample program

---

$Q_1$	0	$R$	$Q_2$
$Q_1$	1	0	$Q_1$
$Q_2$	0	1	$Q_3$
$Q_2$	1	$R$	$Q_2$

- The symbol “R” has a fixed meaning, since it is the instruction to move one square to the right
- But “0” and “1” do not mean anything

# Running the program

$Q_1$	0	<u>1</u>	1	0	1	1	0
$Q_1$	0	<u>0</u>	1	0	1	1	0
$Q_2$	0	0	<u>1</u>	0	1	1	0
$Q_2$	0	0	1	<u>0</u>	1	1	0
$Q_3$	0	0	1	<u>1</u>	1	1	0

# An interpretation function

An interpretation function gives a *semantics*

- assigns objects to symbols

“1”  $\rightarrow$  1

“0”  $\rightarrow$  punctuation mark

- makes it possible to interpret the TM as computing the function of addition

# Syntax tracking semantics

---

## Syntax:

“ $n$ ” = a string of  $n$  “1”s bounded by “0”s

“ $m$ ” = a string of  $m$  “1”s bounded by “0”s

“ $n + m$ ” = a string of  $n + m$  “1”s bounded by “0”s

## Semantics:

“ $n$ ” designates  $n$

“ $m$ ” designates  $m$

## Isomorphism

Given inputs “ $n$ ” and “ $m$ ” the TM outputs “ $n + m$ ”  
just when

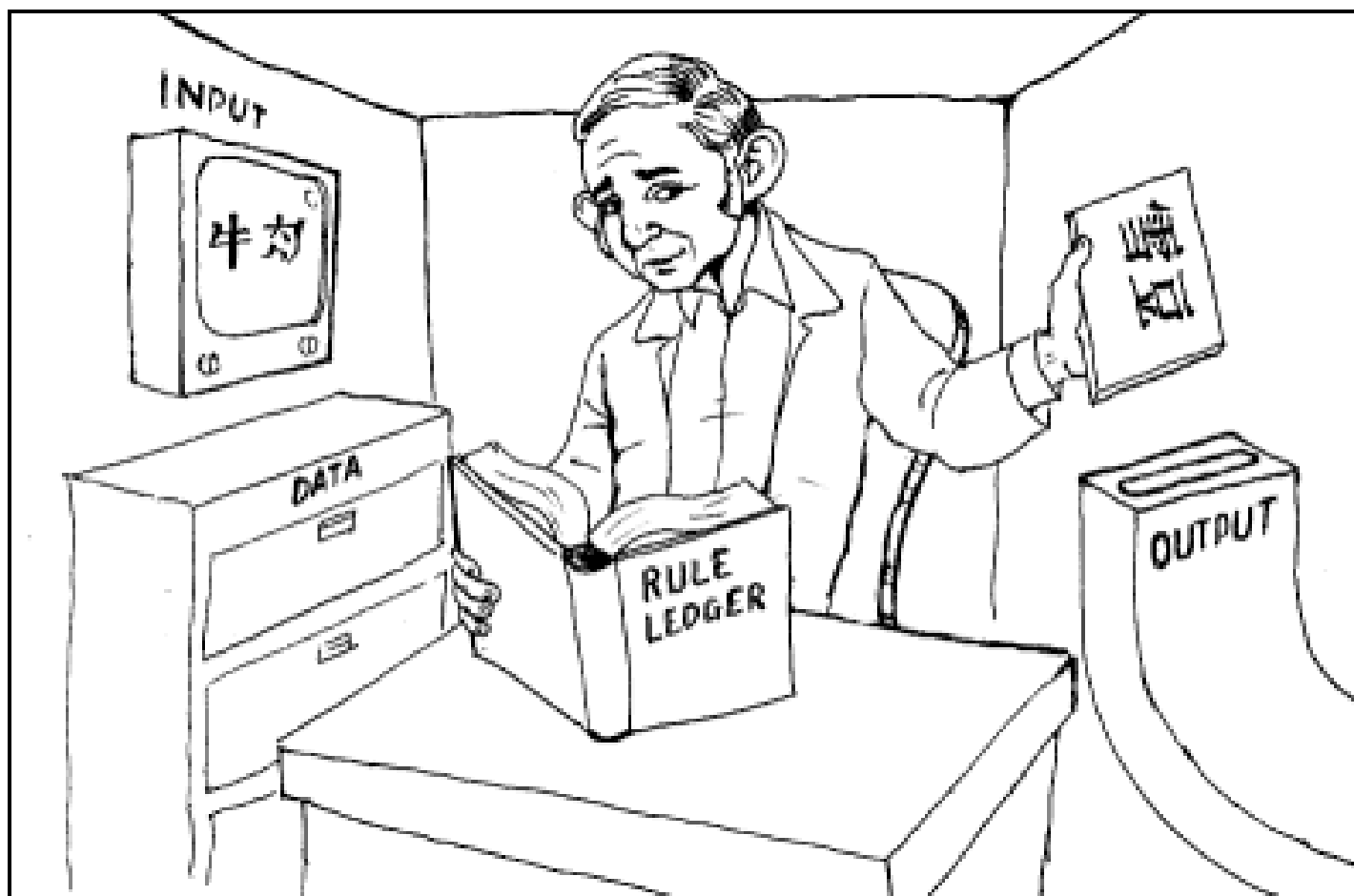
“ $n + m$ ” designates the sum of  $n + m$

# Background to the Russian room

---

- The **Russian room argument** exploits an intuitive contrast between
  - The way that the outputs of a computer result from operations on strings of symbols (“1”s and “0”s)
  - The way that human behavior results from rational thought involving propositional attitudes
- Searle uses the CRA to argue that this contrast is fatal to the project of strong AI (idea that appropriately programmed computers might be minds)
  - PSSH is committed to strong AI

# The Russian room





# The main claims

---

- The Russian Room is input-output identical to a real Russian speaker
- The “internal processing” in the Russian room is purely syntactic (based on the shapes of the symbols)
- The person in the Russian room has no understanding of Russian

Therefore, what is going on in someone who really does understand Russian (or anything else) cannot be the sort of processing that takes place in the Russian room

# What is genuine understanding?

---

- Clearly cannot be understood in purely behavioral terms
  - i.e. producing the appropriate outputs for given inputs
  - The CR passes the Turing Test
- Searle: “Understanding a language, or indeed having mental states at all, involves more than just having a bunch of formal symbols. It involves having an interpretation or a meaning attached to those symbols” (In Chalmers, p. 671)

# Possible responses

---

- (1) Reject the intuition that the CR does not understand Russian
  
- (2) Concede that the CR does not genuinely understand Russian, but find an alternative explanation of the lack of understanding that does not rule out strong AI
  
- (3) Concede that the Russian room does not genuinely understand Russian, but show how we might build up from the CR to a system that does understand Russian

## Strategy 2: system reply

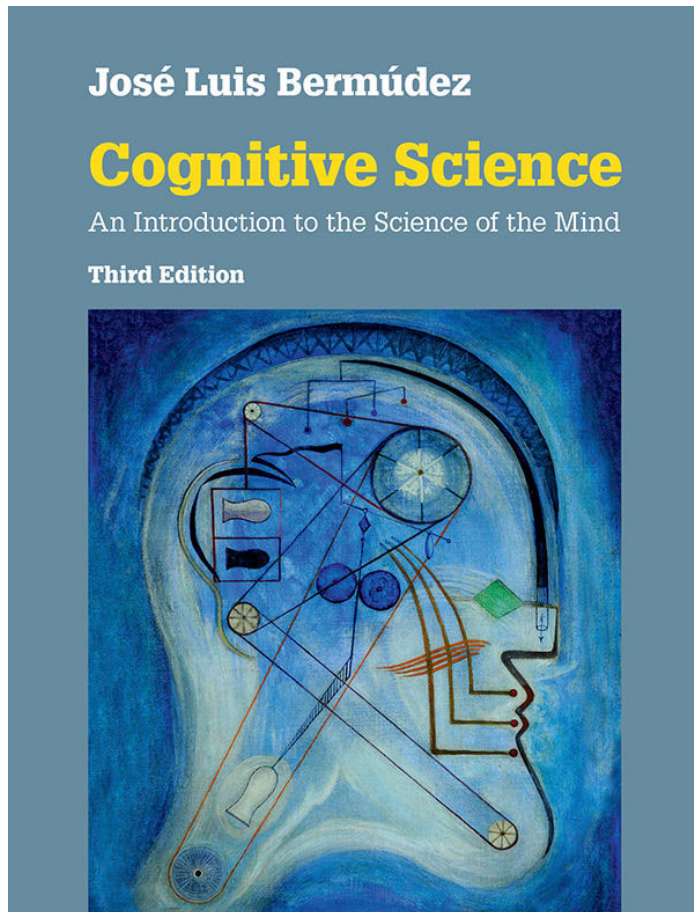
---

- The thought experiment is set up so that the question of whether the CR understands Russian is equivalent to the question of whether the person in the CR understands Russian
  
- But even if we agree that the the person in the room only has a “phrase book” understanding of Russian, this is perfectly compatible with the system as a whole having genuine linguistic understanding


# Strategy 3: robot reply

---

- The input-output test is not a good criterion for genuine understanding
  - It is purely verbal
- A much better test of linguistic understanding is whether the CR can interact with the world appropriately
  - obey instructions and commands
  - name and describe objects correctly
  - initiate conversations in a relevant manner



José Luis Bermúdez,  
**Cognitive Science:**  
**An Introduction to the Science of the Mind,**  
 3<sup>rd</sup> ed., Cambridge University Press, 2020.  
 Chapter 4 (Section 4.3)



## CHAPTER FOUR

# Physical Symbol Systems and the Language of Thought

**OVERVIEW 99**

<p><b>4.1 The Physical Symbol System Hypothesis</b> 100</p> <p>Symbols and Symbol Systems 101</p> <p>Transforming Symbol Structures 102</p> <p>Intelligent Action and the Physical Symbol System 106</p> <p><b>4.2 From Physical Symbol Systems to the Language of Thought</b> 106</p>	<p>Intentional Realism and Causation by Content 108</p> <p>The Language of Thought and the Relation between Syntax and Semantics 110</p> <p><b>4.3 The Russian Room Argument and the Turing Test</b> 114</p> <p>Responding to the Russian Room Argument 117</p>
--	---

### Overview

The analogy between minds and digital computers is one of the most powerful ideas in cognitive science. The physical symbol system hypothesis, proposed in 1975 by the computer scientists Herbert Simon and Allen Newell, articulates the analogy very clearly. It holds that all intelligent behavior essentially involves transforming physical symbols according to rules. Section 4.1 explains the basic idea, while Section 4.2 looks at the version of the physical symbol system hypothesis developed by the philosopher Jerry Fodor. Fodor develops a subtle and sophisticated argument for why symbolic information processing has to take place in a language of thought.

Both the general physical symbol system hypothesis and the language of thought hypothesis distinguish sharply between the syntax of information processing (the physical manipulation of symbol structures) and the semantics of information processing. The philosopher John Searle has developed a famous argument (the Chinese room argument) aiming to show that the project of modeling the mind as a computer is fatally flawed. We look at a version of his argument and at some of the ways of replying to it in Section 4.3.

99