Symbol Manipulation and Intentionality

Minds, Brains, and Programs

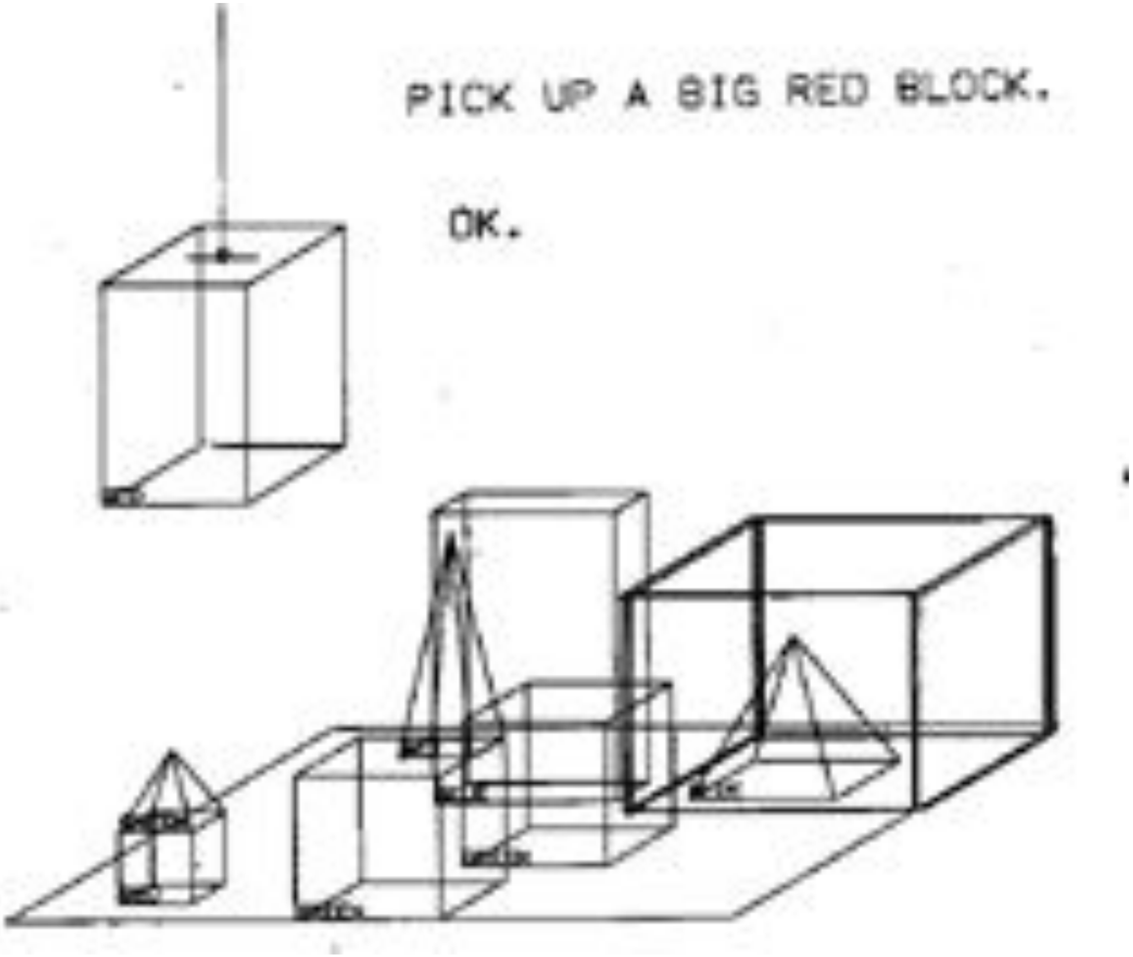
- Searle is addressing a few different but interconnected issues
- One is the relationship between symbol manipulation and *intentionality*
- Another is *functionalism*
- Another is the difference between “strong AI” and “weak AI”
- Another is work that was being done in artificial intelligence around the time the article was published
- (Let’s look at this work first.)

Terry Winograd’s SHRDLU (1971)

Terry Winograd was a grad student at MIT whose dissertation project involved creating a “blocks world” (i.e., a small, digital world of shapes)

A human user, using natural language expressions, could ask the computer about the world and could command the computer to manipulate the world

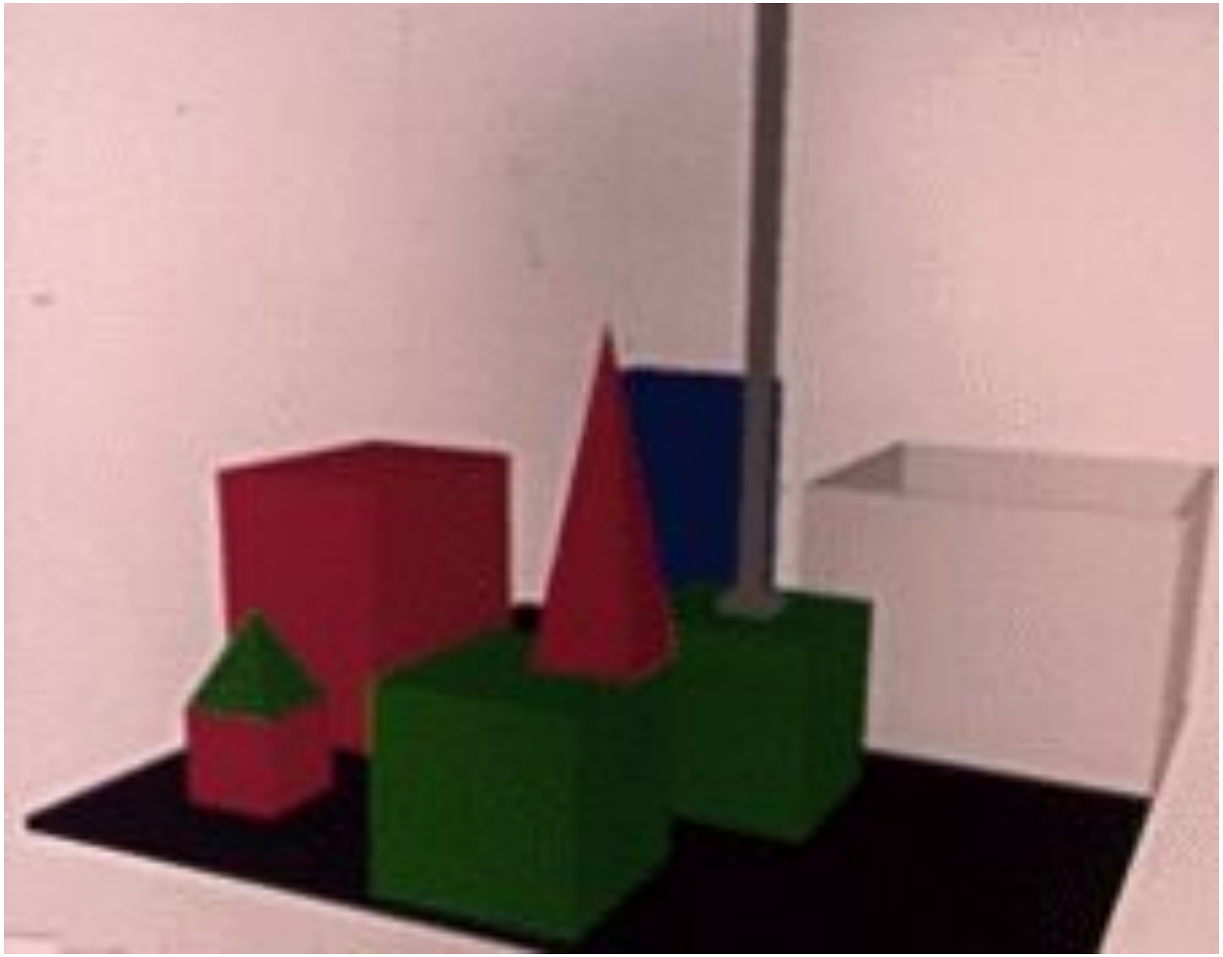
original



PICK UP A BIG RED BLOCK.

OK.

colored rendering



SHRDLU could...

- Describe (when asked) what was in the world (e.g., is there a green triangle on top of a box)
- Move things around (e.g., put a triangle on a box)
- Answer questions about what was possible or not (e.g., can you put a triangle on a triangle?)

conversation

Person: WHICH CUBE IS SITTING ON THE TABLE?

Computer: THE LARGE GREEN ONE WHICH SUPPORTS THE RED PYRAMID.

Person: IS THERE A LARGE BLOCK BEHIND A PYRAMID?

Computer: YES, THREE OF THEM: A LARGE RED ONE, A LARGE GREEN CUBE, AND THE BLUE ONE.

Person: PUT A SMALL ONE ONTO THE GREEN CUBE WHICH SUPPORTS A PYRAMID.

Computer: OK. (does it)

Person: PUT THE LITTLEST PYRAMID ON TOP OF IT.

Computer: OK. (does it)

you can see how predicate logic is useful here

Person: IS THERE A LARGE BLOCK BEHIND A PYRAMID?

To answer the question, you ask whether the following is true: “There is a large block behind a pyramid.”

How would you translate that sentence into predicate logic?

Roger Shank’s work

- Developed a software program that could answer simple questions about a story
- To illustrate the issues, Searle gives the following example:

“A man went into a restaurant and ordered a hamburger. When the hamburger arrived it was burned to a crisp, and the man stormed out of the restaurant angrily, without paying for the hamburger or leaving a tip.”

Did the man eat the hamburger?

We’re not told this information, but in the 1970s there was work on “case-based reasoning” by Roger Schank (and others) that allowed computers to answer questions like this.

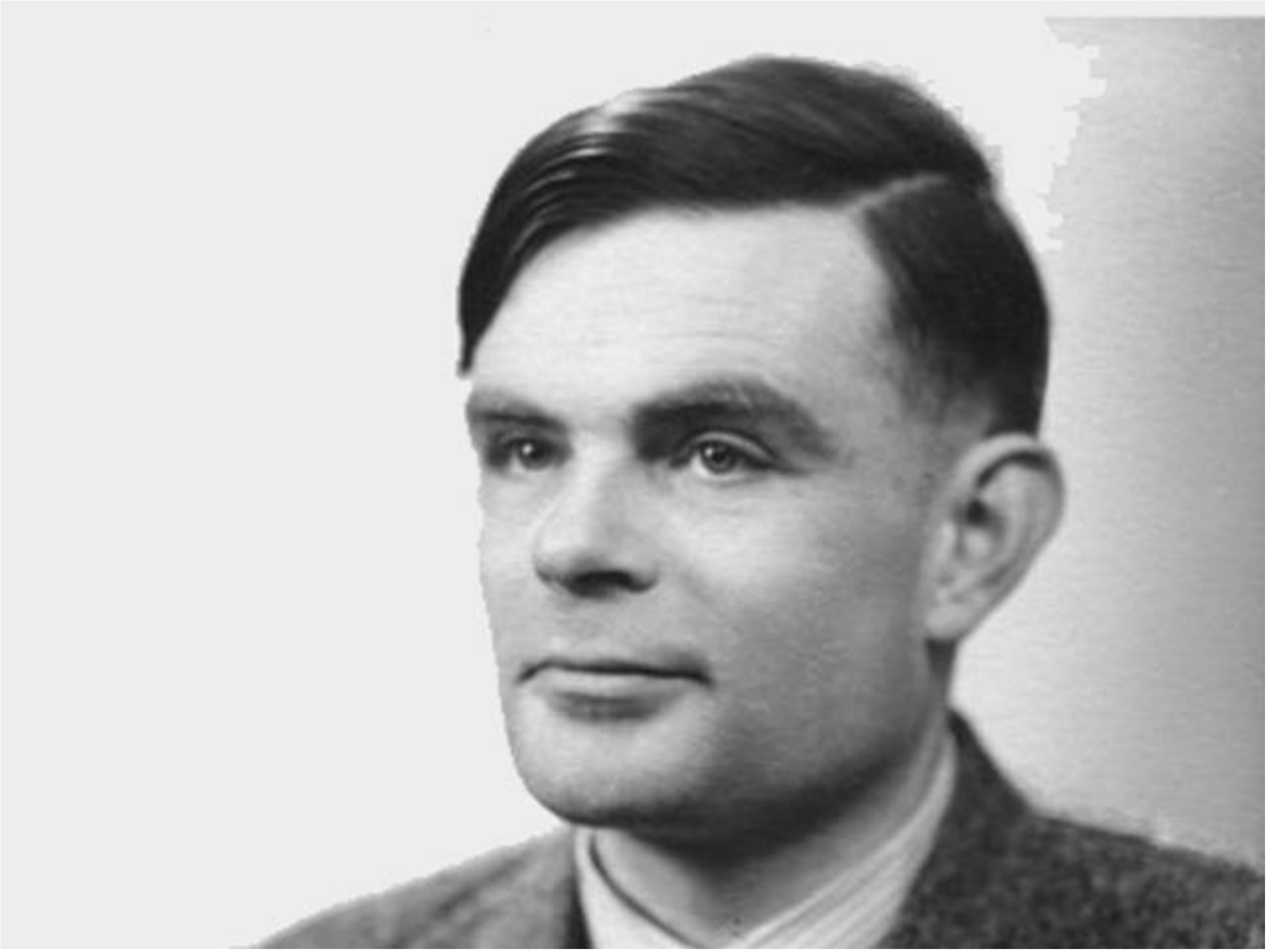
In short, the program “knows” how a sequence of events would naturally unfold, and then might be cued in to deviations from that occurring.

Enthusiasts of AI thought we were starting to build computers that could think and understand.

This is very much related to the Turing Test

Alan Turing (1912-1954)

Alan Turing (1912-1954)



“Computing Machinery and Intelligence” (1950)

Imagine you have one person who is asking questions of both a computer and a real person. (Variant of the “imitation game.”)

If the person asking the questions can’t tell which is the computer and which is the real person, then the computer passes there Turing test

(There are other formulations of the Turing test, but the basic idea is always the same.)

There's been a good deal of discussion about what passing the Turing test would actually accomplish.

Is passing the test really sufficient for saying a machine can think?

There is also the charge of *operationalism* here, that is, to "think" is just to be capable of passing the Turing test

Daniel Dennett on a “great city”

a great city is one in which one can:

- have a nice French meal
- go to the opera
- see a Rembrandt

Dennett’s point is that clearly we should take this as *evidence* for a city’s being great, not a *definition* of a city’s being great

After all, we could have a city with one Rembrandt, one French chef, and one orchestra, but is terrible otherwise (e.g., in the midst of a civil war, disease outbreaks, looting, etc.)

Let’s get to Searle’s Chinese Room

First, what is “intentionality”?

“aboutness”

- Comes from the Latin “intendere”, “to point at”
- E.g., your belief that there’s a staircase outside the door has intentionality because it is about something, namely, the staircase
- Beliefs, desires, and goals have intentionality
- Not all mental states do (e.g., undirected anxiety, depression)

And what is “functionalism”?

In the context of philosophy of mind, functionalism is the hypothesis/view that the materiality of the brain is not essential to cognition

That is, if diodes (or beer cans, water pipes, whatever) were arranged in the correct way, and interacted in the correct way, then they could do everything that a brain does

And the idea here is that a neuron fires (or doesn’t) which influences the probability that some other neuron will fire (or won’t)

So if we replace each neuron in your brain with any object that fired (or didn’t) and influenced whether some other object would fire (or not), then it would replicate the thinking, and experience, of your own brain

Often in cognitive science you here the mind described as “software” and the brain as “hardware”.

A functionalist believes that what the hardware is made out of is unimportant; all that matters is what it does.

functionalism is controversial

- Most cognitive scientists endorse it (I think)
- Most people working on AI do, too
- With philosophers of mind, it’s mixed (e.g., Searle clearly rejects it)

Ned Block’s “Troubles with Functionalism” (1978)

The China Brain Thought Experiment

No relationship to the Chinese Room, aside from being another thought experiment against functionalism

China was chosen because it has the biggest population of any country on earth

This isn’t exactly how Block’s thought experiment worked, but it’s more relevant to the present discussion...

Suppose that China has 100 billion people, and that we give each person a little hand-held machine that beeps when you press a button. And we also hook up electrodes to each neuron in some person’s brain (mine, say) and measure whether that neuron is firing or not over the course of 5 seconds. Over that 5 second period, I will of course have some phenomenal (i.e., subjective) experience. If we instruct the population of China to mimic the firing of my neurons with their little machines, will the nation of China have the same phenomenal experience I had?

Ned Block says clearly it will not. So he believes functionalism is false.

What’s the difference between “strong” and “weak” AI?

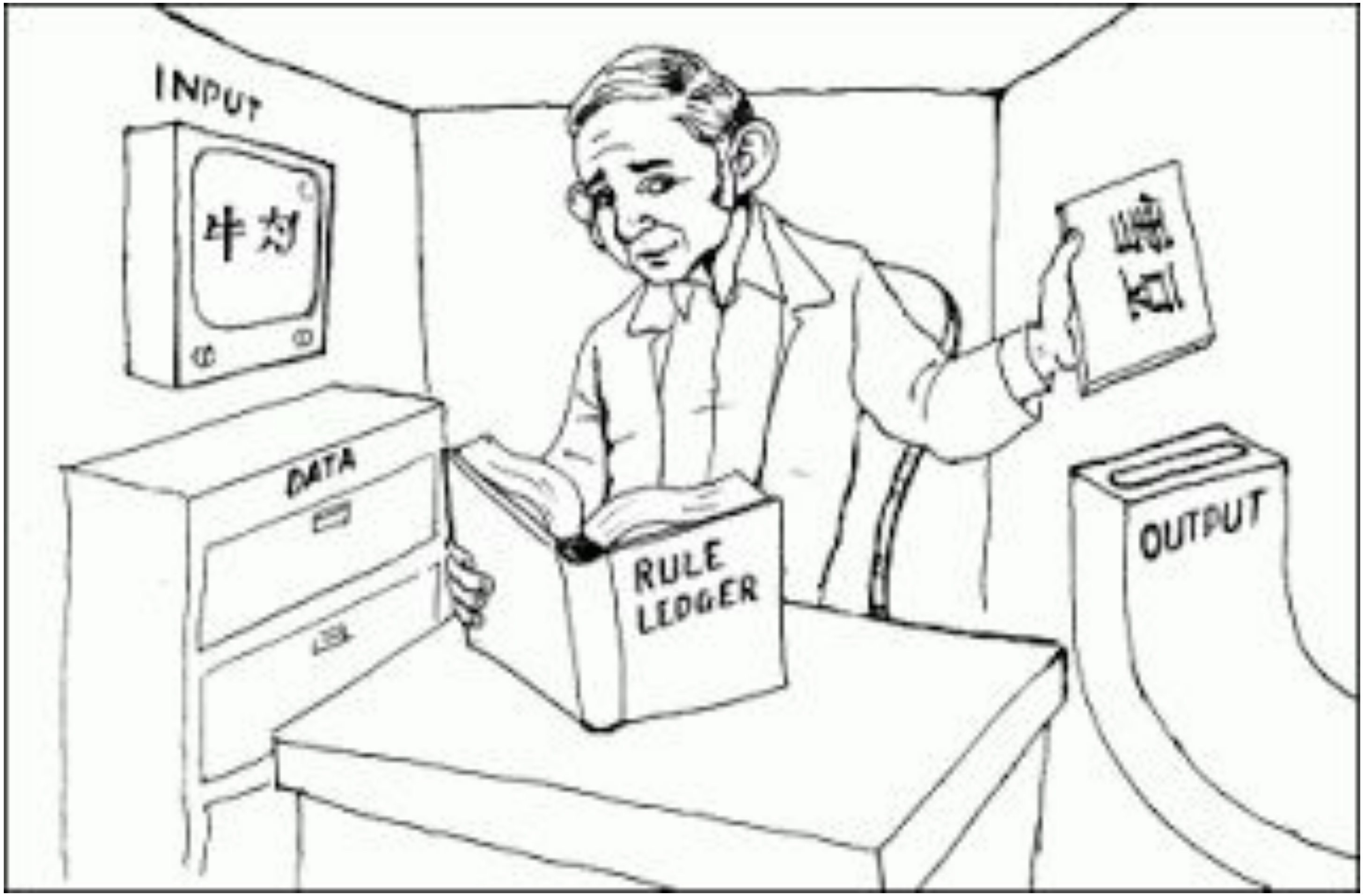
Roughly...

- weak AI just uses computers to help us understand how the mind works
- strong AI is the idea that computers can actually think, understand, or experience in the way that humans can

The Chinese Room

- Searle (who knows no Chinese) is locked in a room
- He has a bunch of Chinese symbols and a rule book (written in English) that tells him how to match certain Chinese symbols with others.
- Someone puts Chinese symbols into the room, he checks the rule book for how to respond, then produces a response.

The Chinese Room



Searle’s Conclusions

(1) Instantiating a computer program is not sufficient for intentionality

After all, Searle instantiates the program, but he does not know what the Chinese symbols mean. For him, the Chinese symbols are not “about” anything.

(In this context, we can really treat “intentionality” and “understanding” as synonymous.)

Searle’s Conclusions

(2) Functionalism is implausible

Functionalism says that what matters are the functional relationships between the parts of a system, not their materiality. But since Searle, by merely by running this program, doesn’t know Chinese, the materiality of the brain must matter.

Searle’s Conclusions

(2*) Strong AI is implausible

Strong AI is predicated on functionalism, and functionalism is implausible, for the reasons given in the previous slide.

Important to note

Searle does *not* say that machines can’t think.

Indeed, he says the brain is a machine and can surely think.

Rather, he says that a machine has to be sufficiently like the brain in order to think.

See p. 422

A few things to note

There’s a reading of Searle (1980) according to which he’s really refuting behaviorism, the idea that what it is to know Chinese (say) is to produce appropriate responses to stimuli.

Or another reading is that Searle (1980) simply shows that passing the Turing Test should not be treated as a sufficient condition for determining whether a machine can think.

the “systems reply”

Perhaps *Searle* doesn’t know Chinese, but “the room” (i.e., the whole “system” does)

Dennett seems to have this response

“Searle observes: ‘No one would suppose that we could produce milk and sugar by running a computer simulation of the formal sequences in lactation and photosynthesis, but where the mind is concerned many people are willing to believe in such a miracle.’ I don't think this is just a curious illustration of Searle's vision; I think it vividly expresses the feature that most radically distinguishes his view from the prevailing winds of doctrine. For Searle, intentionality is rather like a wonderful substance secreted by the brain the way the pancreas secretes insulin. Brains produce intentionality, he says, whereas other objects, such as computer programs, do not, even if they happen to be designed to mimic the input-output behavior of (some) brain.”

Dennett goes on...

“[Searle] can't really view intentionality as a marvelous mental fluid, so what is he trying to get at? I think his concern with internal properties of control systems is a misconceived attempt to capture the interior point of view of a conscious agent. He does not see how any mere computer, chopping away at a formal program, could harbor such a point of view. But that is because he is looking too deep. It is just as mysterious if we peer into the synapse-filled jungles of the brain and wonder where consciousness is hiding. It is not at that level of description that a proper subject of consciousness will be found. That is the systems reply, which Searle does not yet see to be a step in the right direction away from his updated version of elan vital”.

How does Searle respond?

“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” (419).

the “robot reply”

"Suppose we wrote a different kind of program from Schank's program. Suppose we put a computer inside a robot, and this computer would not just take in formal symbols as input and give out formal symbols as output, but rather would actually operate the robot in such a way that the robot does something very much like perceiving, walking, moving about, hammering nails, eating, drinking - anything you like. The robot would, for example, have a television camera attached to it that enabled it to 'see,’ it would have arms and legs that enabled it to 'act,' and all of this would be controlled by its computer 'brain.' Such a robot would, unlike Schank's computer, have genuine understanding and other mental states” (p. 420)

How does Searle respond?

But the answer to the robot reply is that the addition of such "perceptual" and "motor" capacities adds nothing by way of understanding, in particular, or intentionality, in general, to Schank's original program. To see this, notice that the same thought experiment applies to the robot case. Suppose that instead of the computer inside the robot, you put me inside the room and, as in the original Chinese case, you give me more Chinese symbols with more instructions in English for matching Chinese symbols to Chinese symbols and feeding back Chinese symbols to the outside. Suppose, unknown to me, some of the Chinese symbols that come to me come from a television camera attached to the robot and other Chinese symbols that I am giving out serve to make the motors inside the robot move the robot's legs or arms. It is important to emphasize that all I am doing is manipulating formal symbols: I know none of these other facts. I am receiving "information" from the robot’s "perceptual" apparatus, and I am giving out "instructions" to its motor apparatus without knowing either of these facts. I am the robot's homunculus, but unlike the traditional homunculus, I don't know what's going on.

So does the Chinese Room thought experiment show that a computer could never “understand” human language?

Pair off into groups of 3-4, and comes up with possible responses to Searle (even if you agree with Searle!)

That is, come up with possible reasons why the thought experiment fails to show that a computer can never understand language.