24.09 Minds and Machines
Fall 11 HASS-D CI

the Chinese room argument contd., and lessons from it

the Turing test (if we have time—er, we didn't)

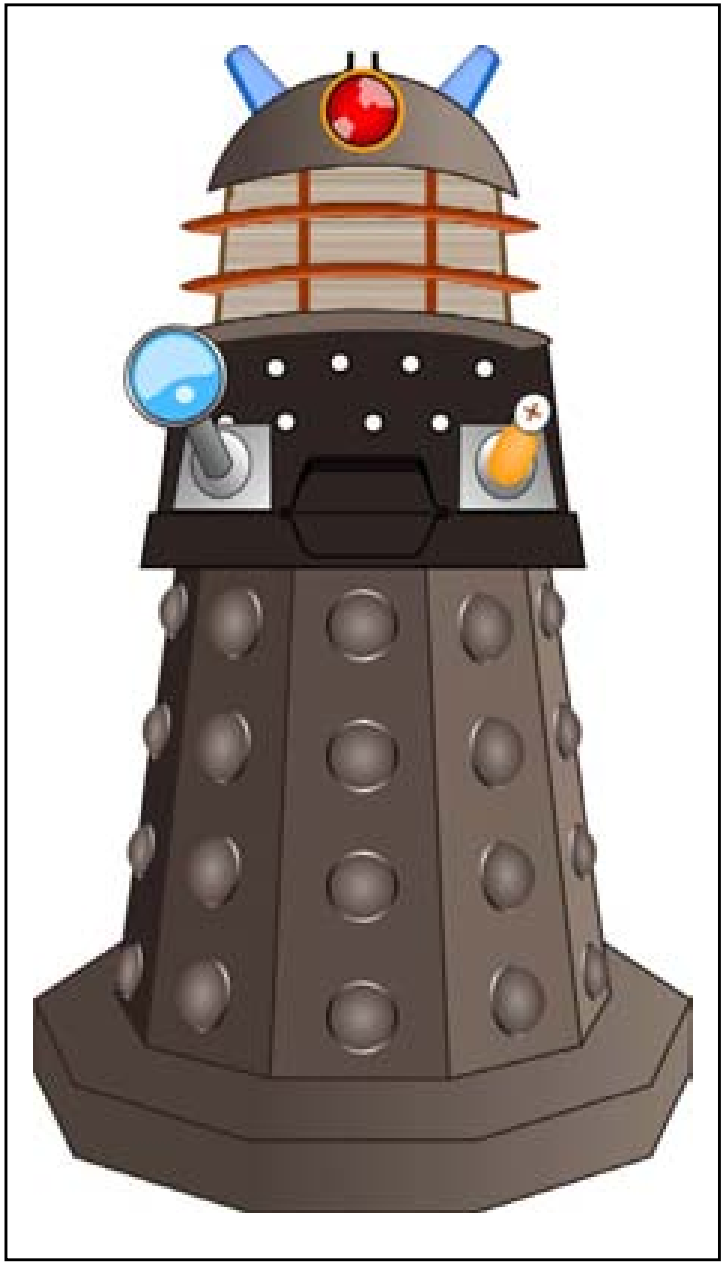

Image by MIT OpenCourseWare.

Strong AI

STRONG AI: an appropriately programmed computer literally has mental states (in particular, cognitive states)

more precisely: there is a computer program P such that, *necessarily*, any computer running P is in such-and-such mental states (believes it's raining in Beijing/intends to vote for Perry/wants Romney to win/....)

programs

a program: an algorithm (mechanical recipe) for transforming symbols into symbols

the thought experiment exploits the fact that computer programs can be 'multiply realized'

that is, computer programs can be implemented on a diverse range of hardware

in particular, the Chinese room

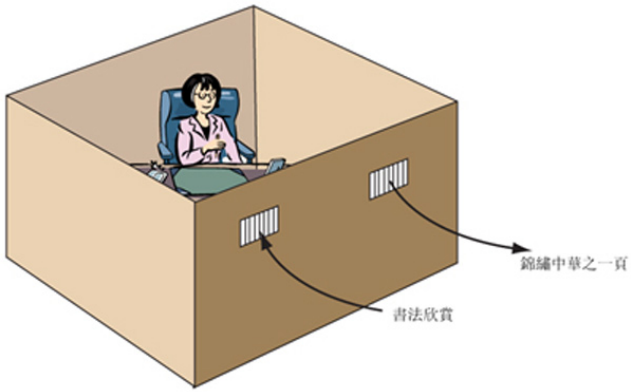

Image by MIT OpenCourseWare.

Turing on multiple realizability

Importance is often attached to the fact that modern digital computers are electrical, and that the nervous system also is electrical. Since Babbage's machine was not electrical, and since all digital computers are in a sense equivalent, we see that this use of electricity cannot be of theoretical importance...The feature of using electricity is thus seen to be only a very superficial similarity. If we wish to find such similarities we should took rather for mathematical analogies of function. (from 'Computing machinery and intelligence')

Searle's argument

…you behave exactly as if you understood Chinese, but all the same you don't understand a word of Chinese. But if going through the appropriate computer program for understanding Chinese is not enough to give you an understanding of Chinese, then it is not enough to give any other digital computer an understanding of Chinese.

so, strong AI is false ?

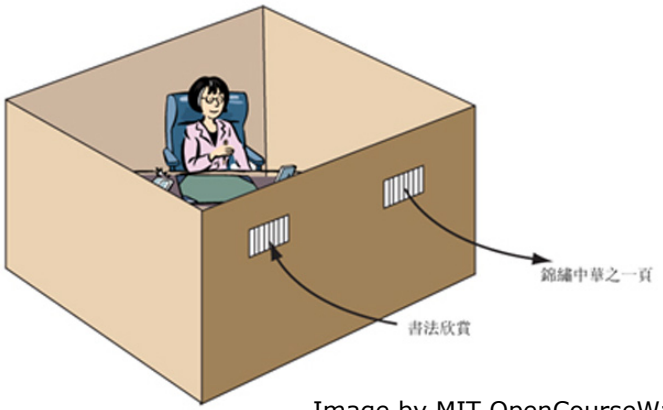

Image by MIT OpenCourseWare.

Searle's argument...

… appears to rely on the mistaken principle that if x is part of y, and x isn't F, then y isn't F. (My liver is part of me, and doesn't teach philosophy, but that doesn't mean I don't.)

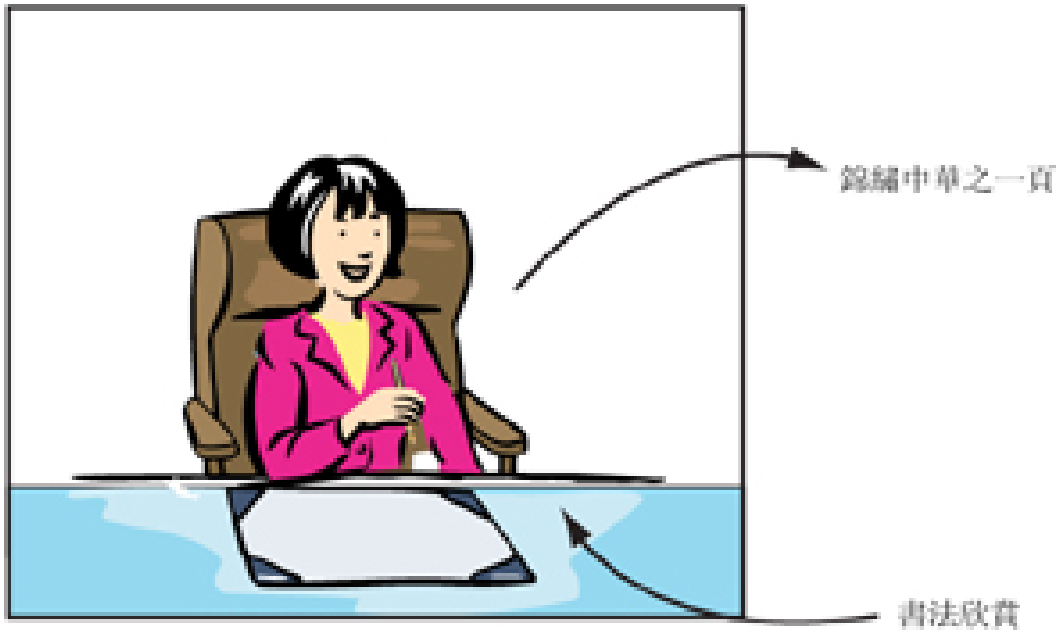

Image by MIT OpenCourseWare.

digression on Alan Turing and Turing machines

Image removed due to copyright restrictions.
A photograph of Alan Turing (1912 - 54).

wartime codebreaker, founder of computability theory

invented 'Turing machines'

also invented the 'Turing test' (of which more later)

Image removed due to copyright restrictions. To view the story "PM Apology After Turing Petition", please go to the website: http://news.bbc.co.uk/2/hi/8249792.stm.

a Turing machine

states: S1, S2,…,Sn

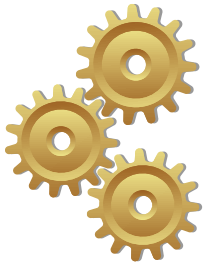

Image by MIT OpenCourseWare.

head

{R, L, 1, 0, halt}

…1000111110000111110111000000000…

tape

a simple Turing machine table

state scanned cell	S1	S2
1	R, S1	H
0	1, S2	H

the machine in action

0000000111110000000000000

S1 S2 H

state / scanned cell	S1	S2
1	R, S1	H
0	1, S2	H

computable functions

a function from the natural numbers to the natural numbers is Turing-computable iff ("if and only if") some Turing machine computes it

Church-Turing thesis: every computable function is Turing-computable

let $f(x) = 1$ if neutrinos can travel fast than light, and 0 otherwise

is the function f Turing-computable?

the systems reply

the whole system understands Chinese, not Searle

this isn't really a 'reply'–it's the thesis that Searle is supposed to be refuting

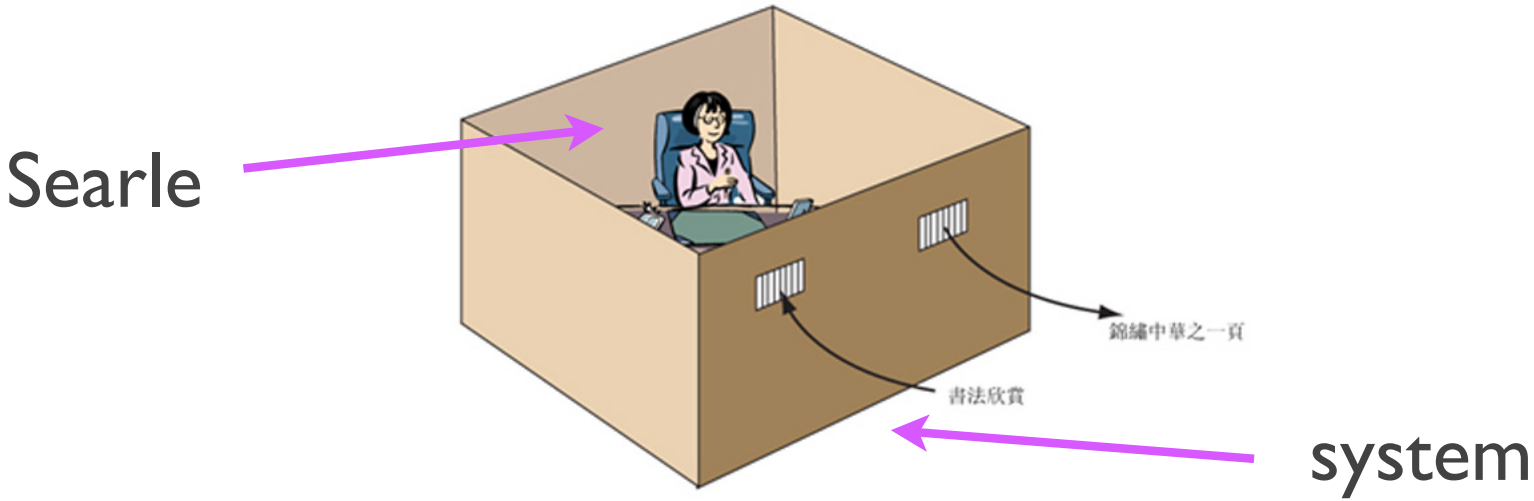

Image by MIT OpenCourseWare.

Searle's reply...

…is quite simple: Let the individual internalize all of these elements of the system...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.

Searle, 'Minds, Brains, and Programs'

Searle (memorizes instructions)

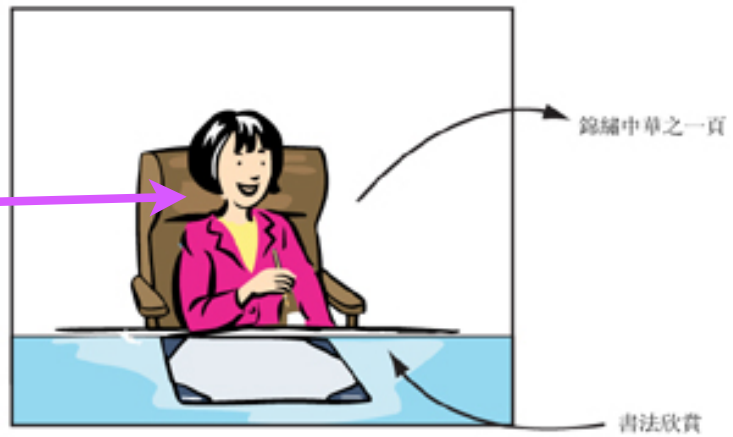

Image by MIT OpenCourseWare.

Searle's argument...

… appears to rely on the mistaken principle that if x is part of y, and y isn't F, then x isn't F. (My liver is part of me, and I don't weigh 1 pound, but that doesn't mean my liver doesn't.)

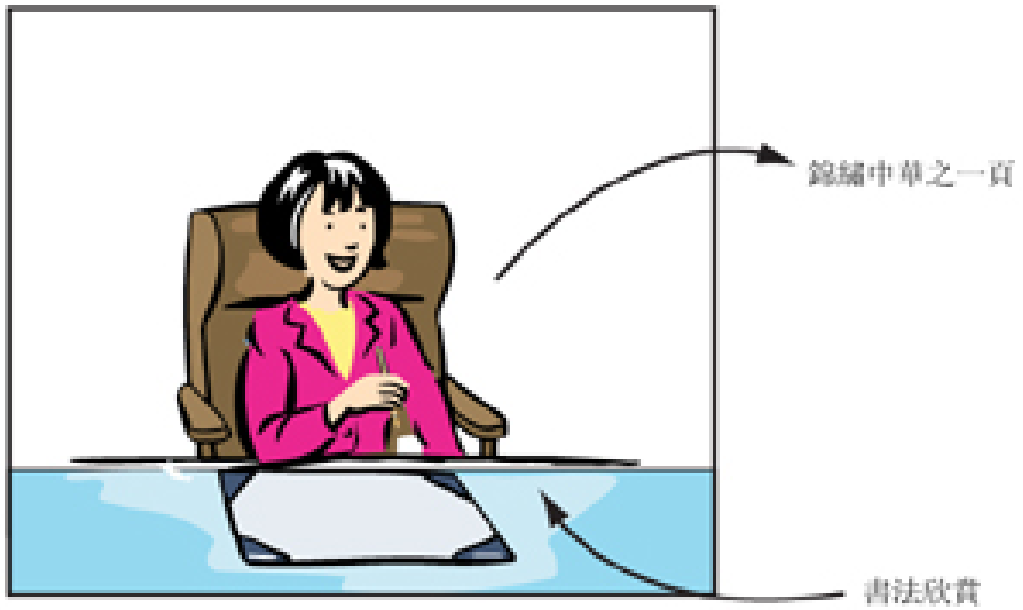

Image by MIT OpenCourseWare.

the robot reply

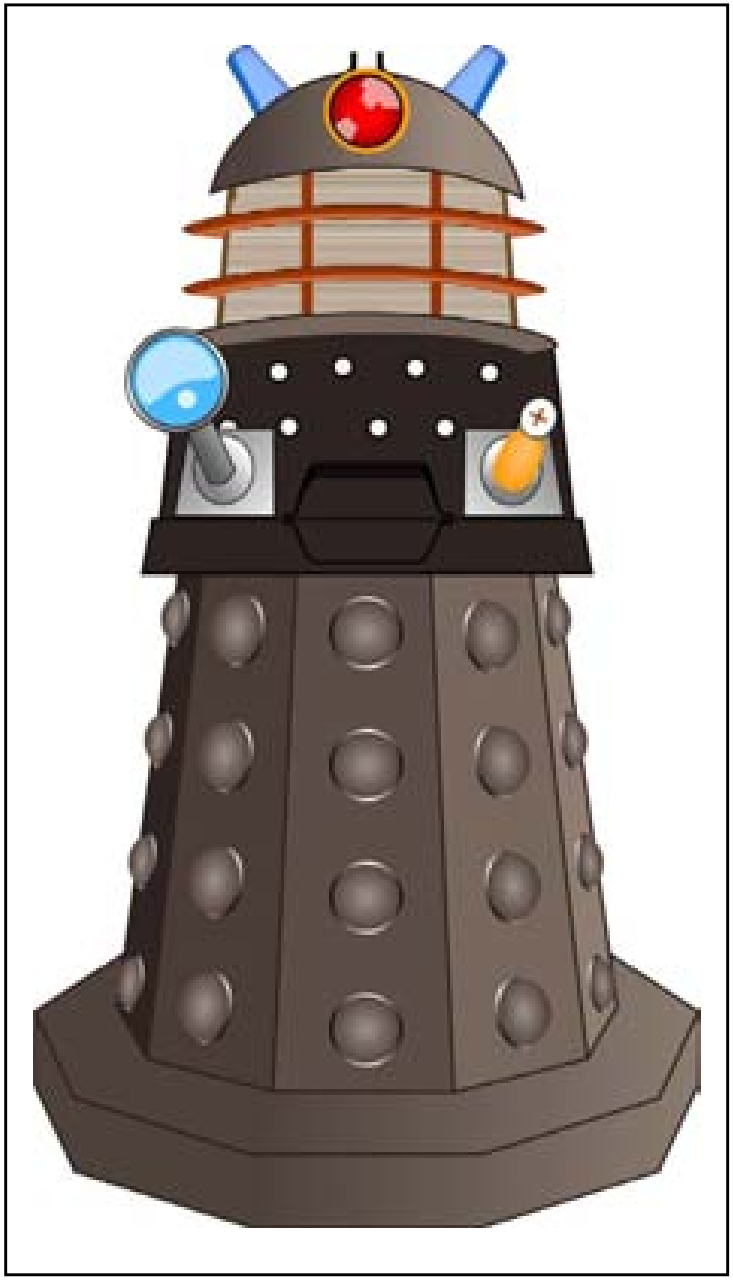

Image by MIT OpenCourseWare.

Inside a room in the robot's skull I shuffle symbols…As long as all I have is a formal computer program, I have no way of attaching any meaning to any of the symbols. And the fact that the robot is engaged in causal interaction with the outside world won't help me…

strong strong vs. weak strong AI

STRONG STRONG AI: there is a computer program (i.e. an algorithm for manipulating symbols) such that any (possible) computer running this program literally has cognitive states

WEAK STRONG AI: there is a computer program such that any (possible) computer running this program and embedded in the world in certain ways (e.g. certain causal connections hold between its internal states and states of its environment) literally has cognitive states

morals from the Chinese room

Searle's official argument against strong AI fails

but he does have a point, namely that merely implementing a program is arguably insufficient for (underived) intentionality

something else is needed—perhaps certain kinds of causal connections between the system and its environment

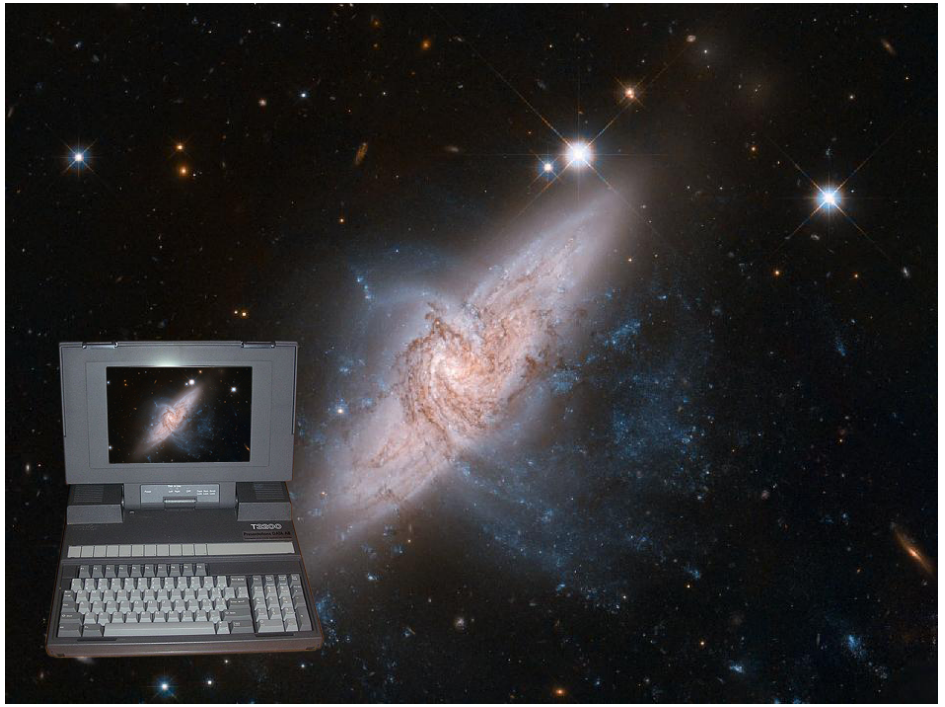

Image by MIT OpenCourseWare. Public domain photo courtesy of NASA.

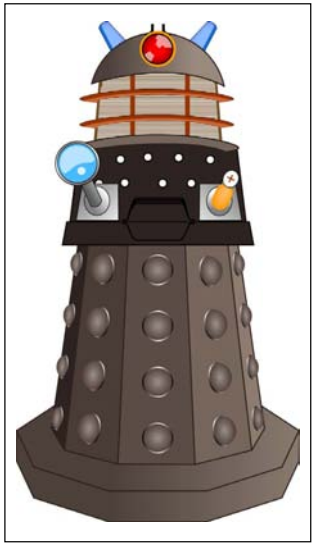

Image by MIT OpenCourseWare.

24.09 Minds and Machines

Fall 2011