24.119 Minds and Machines

Handout 1: Searle's Chinese Room Argument

Intentionality: that feature, possessed by (e.g.) words and mental states, of being "about" -- representing, referring to -- something. The belief that Fido is furry is a mental state that is about Fido. And the word `Fido' refers to the dog, Fido. So these are two examples of intentionality. (NB: don't confuse intentionality with intending something -- the latter is just another example (along with believing and desiring) of an intentional mental state.) [See Block, The Mind as..., sect. 2; Byrne, Intentionality]

Something has *derived intentionality* just in case it has intentionality in virtue of the intentionality of something else. Plausibly, `dog' refers to dogs in virtue of the beliefs, intentions, etc., of English speakers -- hence `dog' has derived intentionality. My belief that dogs have fur is an intentional mental state, and doesn't have its intentionality in virtue of the intentionality of anything else -- hence my belief has *underived* (or *original*) intentionality. If thinking is conducted in a language written in the brain, then the words of this language have underived intentionality. [See Block, sect. 2]

Searle's Chinese Room argument is directed against the claim that instantiating a computer program is sufficient for *underived intentionality*.

WEAK AI
The principle value of the computer in the study of the mind is that it gives us a very powerful tool -- e.g. it enables us to *simulate* various kinds of mental processes. (Cf. WEAK ARTIFICIAL METEOROLOGY.)

Obviously correct (ditto WEAK AM).

STRONG AI
An appropriately programmed computer literally has cognitive states. (Cf. STRONG AM -- an appropriately programmed computer literally has meteorological states.)

Disputable, and disputed by Searle. (STRONG AM, at least, is obviously false.)

Let us distinguish two versions of 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.

As is evident from Searle's reply to the "Robot Reply" (p. 672), he takes the Chinese Room argument to work equally well against both STRONG STRONG AI and WEAK STRONG AI. So, does it?

Now, the proponent of either thesis should not say that Searle himself in the Chinese Room understands Chinese -- Searle is not the computer, but only part of its machinery. So the right reply is the Systems Reply. (Calling it a "reply" is misleading: it is the thesis that is up for refutation in the first place.) And Searle's response to that "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" (minds, brains, and programs). But this 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 maybe my liver does.) [See Block, sect. 4]

So the Chinese Room argument seems quite impotent against either thesis: the argument makes essential use of the premise that Searle, in various situations, does not understand Chinese, but this is quite irrelevant. (See, however, Searle's Chinese Room MITECS entry.)

Nonetheless, is either thesis *true*?

One might well doubt STRONG STRONG AI, on the grounds that nothing could make a symbol (in some computer) refer to a thing -- our fair city of Cambridge, for example -- if the computer has forever been floating off in deep space, causally isolated from the thing.

As to WEAK STRONG AI, there is a large literature on what sorts of connections between symbols and other parts of the world would suffice to give those symbols underived intentionality. (See Block, sect 3; Chalmers, pp. 473-4.) This issue will surface again later.
