Brette Fitzgibbon

a) Summarize the Chinese room argument.

John Searle imagines that he is in a “Chinese room”, a room with a seemingly infinite number of volumes that together comprise a dictionary of questions written in Chinese characters and their responses in Chinese characters. Chinese speakers pass their questions in through a slot. He finds their question in the dictionary, writes the corresponding response on a slip of paper, and passes it back to the speaker through the slot. He does not understand Chinese, but the Chinese speaker assumes he does because his answers make sense and sound like things a native speaker would say.

b) What does this have to do with computers?

The spirit of the Chinese room argument is how large language models work. Instead of a dictionary, they tokenize input and use linear algebra, calculus, and statistics to put together an appropriate response. While the Chinese room argument sounds a bit far-fetched (how large is this room? How many volumes are there? How does Searle have time to search through them all?), LLMs have enough training data and computing power to interpret the input and formulate a response in less than a second, making a mapping of any question to any response possible without understanding the meaning of anything that is said.

c) Why does Searle believe that it shows that a computer can pass the Turing test without understanding?

Searle defines understanding as knowing the definitions of the words and their connotations when put together. Reading “It’s raining today” and imagining a sky blanketed with deep gray clouds and a torrent of cold water pouring down on people who are trying their best to stay comfortable and dry as they hurry from one indoor location to another, and advising to pack an umbrella to stay as dry as one can. He does not consider what the computer does – mapping “it’s + raining + today” with “pack + an + umbrella” – understanding. “Pack an umbrella” would pass the Turing test because it is an identical response to that borne of Searle’s idea of understanding, without actually using Searle’s idea of understanding.

d) Do you find Searle’s argument convincing? Or do you find one of the responses more appealing?

I support Searle’s argument. I think there is value in understanding what words mean and how they relate to the world. I find the Brain Simulator Reply a little bit annoying because different parts of the brain are triggered when reading something you understand versus matching up seemingly arbitrary images (which is all that the Chinese characters are to the person in the Chinese room) with other seemingly arbitrary images. Not all brain functions are the same. Yet the Robot Reply calls into question what it means to “understand” something, and makes me wonder if Searle and those of us who support his argument might be acting as gatekeepers, excluding AI from the status of Intelligent Being by holding it to a standard that only a human could possess. Who’s to say that my brooding clouds and sheets of rain are better than “it’s + raining + today -> pack + an + umbrella” if the response is identical and therefore equally helpful?