Minds, Brains, and Maybe Programs. And also elephants. Caleb Belth

(This was written for a Philosophy of Mind course at Purdue University in March 2018. I have corrected some grammar since.)

A few decades ago (in 1980 to be precise), the eminent philosopher John Searle proposed a thought experiment attempting to refute the contemporary view that “a system has a mind when the system has a suitable *functional organization*” (Rescorla, 2015, p. 5). This thought experiment, which has become known as the “Chinese Room” has become so prominent that when I find myself talking philosophy of mind with a new acquaintance, which I find myself doing more and more frequently, I find that the acquaintance may bring up what they might call “The Chinese Room paradox,” or something along those lines, which is not quite what Searle suggested, but appears to be some colloquial variant story. However, despite the ubiquity of Searle’s refutations of functionalism and the consequence that machines can’t have minds, the increasing sense of folks is that machines can most certainly have minds. Talk of reaching “Singularity” and anthropomorphic computer talk seems to fill more and more conversations. At the same time, researchers in Computer Science often seem too caught up in engineering new methods of making advertisements effective to spend any time engaging with philosophers. However, I think it is important for CS researchers, which includes myself, to engage in philosophical dialogue. As Noam Chomsky puts it for scientists in general, “It should trouble us that we aren’t thinking about what we’re up to” (Chomsky, 2009). Thus, in this essay, I will attempt to show that Searle’s Chinese Room thought experiment rests upon what I think is a fallacious assumption that symbol manipulation *cannot have meaning*. I will attempt to show that the way in which Searle describes computation in his paper is loaded in such a way that the reader’s intuitions are softened to make them agree that Searle is obviously right in this assumption, but that a system of *meaningful symbol manipulation* can actually be described. It is important to note that I will *not* argue that current research in Artificial Intelligence is doing much of anything remotely similar to what the human brain does. This is not an appeal to the wonders of AI. Indeed, I think terms like “Artificial Intelligence” and “Machine Learning” and “Neural Networks” are horrendously misleading and perhaps even misnomers. I am a proponent of companies and researchers being more clear and precise about what they are *actually* doing. Rather, this essay will appeal to current research and describe the meaning of the symbol manipulations involved, even though that meaning is currently quite thin.

Searle (1980) begins his article with two propositions. From these two propositions, he draws three conclusions. The first proposition is that intentionality in humans and animals is caused by features of the brain. Searle does not discuss this in his article, and neither will I. He states the primary purpose of his article as establishing the validity of his second proposition. Namely, “Instantiating a computer program is never by itself a sufficient condition of intentionality” (p. 1). Searle describes the Chinese Room in order to defend this proposition. The proposition depends upon an assumption which Searle touches on throughout the article. Namely, that instantiation of a computer program is symbol manipulation and that this symbol manipulation is meaningless. In this essay, I will attempt to show that this proposition may not be true by showing that symbol manipulation *can have meaning*. The conclusions from these two propositions are that the instantiation of a computer program cannot explain how the brain produces intentionality, anything that can produce intentionality (like that of a human or animal) must be causally equivalent to the brain, and lastly, for strong AI to succeed, it would need to be able to match the causality of the brain. The first and third conclusions rest, in part, upon the second proposition. Thus, if the assumption upon which proposition two is based is wrong, the proposition is false and the first and third conclusions are no longer necessary.

Now I will describe the Chinese Room Argument. The first, and perhaps most important, point is that Searle does not know Chinese. In this grand experiment, Searle is locked into a room. At his disposal, he has his utterly non-existent understanding of Chinese, his much more existent (and indeed, as a philosopher, probably above average) understanding of English, three sets of Chinese writing, and a set of English rules. The three sets of Chinese writing are a "script," a "story," and "questions" respectively. Searle does not know that the terms ("script", "story", and "questions") are used to refer to these sets of writing; nor does he know that the term "program" is used to refer to the English rules at his disposal (Searle, 1980). The English rules show Searle how to relate questions with stories and likewise relate stories with the script. This scenario, as designed by Searle, mimics what Roger Shank and other researchers at Yale attempted to create a computer program to do—namely to infer common sense responses to questions based on stories in natural language (Schank & Abelson, 1977). The responses were supposed to be sensible enough inferences that an onlooker would think the machine understood the stories. Thus, the "script" in Searle's setup is a set of canned language describing common situations in the Shank setup. The "story" and "questions" in Searle's setup are obviously the story and the questions in Shank's setup. The English rules that Searle has correspond to the computer code in Shank's program.

Searle's claim (1980), as stated before, is that "Instantiating a computer program is never by itself a sufficient condition of intentionality" (p. 1). He argues this proposition by pointing out that, in the case of the Chinese Room, if a computer program could be by itself a sufficient condition for understanding Chinese, then when he runs that program as someone who doesn't know Chinese, it would need to result in him understanding Chinese. However, he does not understand Chinese even while running the program because the program is merely meaningless symbol manipulation. Searle's claim could be restated: no computer program is by itself sufficient for intentionality. Thus, to show this claim to be false, all one needs to do is show a counterexample. Furthermore, the counterexample need only be plausible. This is due to the fact that, if a program that on its own is sufficient for intentionality *could* exist, then it is false to say, "no program on its own is sufficient for intentionality." Searle, in his setup of the Chinese Room, has provided a description of a program, as symbol manipulation, that is supposedly broad enough to stand in place of any computer program. Thus, I will attempt to show that it is not actually broad enough to stand in place of any computer program because the symbol manipulation he describes is *meaningless* while we can (and I will hopefully do so) describe symbol manipulation that is *meaningful.* Thus, if Searle, as someone who doesn't understand Chinese, were to run this program of meaningful symbol manipulation, he would understand at least something of Chinese.

It is useful to note that the Chinese Room seems to be a representation of a Turing Machine program. That is, it represents what computation, under the model of the Turing Machine, is doing. This is a consequence of the fact that Searle (1980) refers to it as "symbol manipulation" throughout the article. Turing computation is not the only model of computation that exists, but it is, at least in a basic sense, the model on which modern computers are built (Rescorla, 2015, p. 3). This is also significant in terms of the history of Functionalism. Hillary Putnam (2002), in *The Nature of Mental States,* uses a variant of Turing Machines to introduce his version of Functionalism. Thus, as Searle attempts to refute functionalism, it makes sense for him to describe a system in line with the one described by Putnam. In this essay, I will describe some programs at a more abstract level. This is not to say that they are more than symbol manipulation. Indeed, modern AI programs are run on modern computers, which, as mentioned above, are roughly Turing Machines. Thus, if we can conceive of, for example, a machine learning program that could be by itself sufficient for intentionality, then this program would be a sufficient counterexample to Searle's proposition, and, as implemented on a modern machine, would still constitute symbol manipulation.

Let us first consider an important historical note from Hillary Putnam's paper in which he introduced his version of functionalism. Putnam uses Turing computation in his description of functionalism. However, it differs in one regard which I think is quite significant. Namely, it introduces stochastic state transitions (Putnam 2002). In contrast to traditional Turing Machines, which move from state to state deterministically as designated by a transition table, Putnam's description uses transitions between states that occur *probabilistically*. Thus, rather than rules of the form "given state S_1 and input I_1, output O_1 and proceed to state S_2," we would have rules of the form "given state S_1 and input I_1, output O_1 and proceed to state S_2 with probability *p* or output O_1 and proceed to state S_3 with probability *q*, or output O_1 and proceed to state S_4 with probability *r*, etc." I can hopefully make it apparent why this is significant with an example. Consider the Chinese Room. According to Searle (1980), when he sees "squiggle, squiggle" he writes "squoggle, squoggle" based on some English rule (i.e. a transition table). Thus, although not explicitly stated as such, it seems that Searle has in mind the traditional, deterministic transition table. A problem with this is that if someone were interrogating the Chinese Room, she could easily determine that the system does not understand the symbols because she could ask the same question over and over again and the room, assuming it is in the same start state[1], would respond with the same answer every time, deterministically. For example, the interrogator

[1] The output and next state are specified in the transition table based on the current state and the current input. Thus, the same query could lead to different outputs if the input state were different each time. Thus, you could potentially create some sort of dynamic response even with a deterministic transition table as long as you had a diverse enough set of input, current state pairs. However, this is much more implausible because the amount of space the transition table would take would blow up much more quickly than with transition probabilities. I don't have any source on this concept, but I developed it after reading the Putnam (2002) article I reference above. Putnam seems to emphasize the probabilistic nature of the Turing Machine and I suspect he would agree on my intuition that a natural language generating system that is deterministic would be so rigid as to be implausible. Hence, stochasticity is necessary for any plausible system description.

might ask "What's wrong?" many times over. The room might deterministically respond "Nothing is wrong" every time. By introducing transition probabilities, the response might differ each time she asked the question. In this scenario, prompted by the same query, the room might respond "Nothing is wrong" a few times in a row, then respond with the troubles of its life, as if it understood that the integrator was interested in listening to its troubles[2]. If the response were random, the interrogator would still obviously know that the room doesn't understand. It is ridiculous to think that random symbol manipulation could constitute understanding any more than deterministic symbol manipulation could. However, the transition probabilities broaden the possibility of dynamic responses. Surely even our intuitions will soften to the possibility of a system that responds dynamically is more plausibly *understanding* our questions than one that responds the same way every time we query it. Although our intuitions softening is only epistemic (i.e. it's not saying anything about the metaphysical *existence* of understanding), all I am looking for by appealing to probability is plausibility because, as I stated earlier, symbol manipulations which could *plausibly* have meaning suffices as a counterexample to the claim that no instantiated program can constitute intentionality. Dynamic symbol manipulation *more plausibly* has meaning than does deterministic symbol manipulation.

Having discussed probability as the first foundational point in finding a counterexample to Searle's proposition, I now want to move on to the second: context. Someone once told me that when interpreting the Bible, "context is king." That statement rings true in all language. Understanding language depends upon context. If I were to say, "An elephant is an animal with large, floppy ears, a long trunk, and which lives in Africa," the word "elephant" here clearly

[2] This is supposed to depict two people talking. One person is genuinely concerned about the other. The later person doesn't understand that the former is genuinely concerned and brushes off their problems with "Nothing is wrong." The later person then picks up from the former's persistence that they are genuinely concerned. Thus, the symbols "What's wrong?" go from meaningless to having the meaning of genuine concern.

means the conceptual *definition* of an elephant. If I were to, on the other hand, say “The elephant was shot by the poacher,” the word “elephant” would here mean a *specific* elephant shot by a specific poacher. To stress the point even more, consider a young child at the zoo. He has just seen the elephants and has moved on to look at the zebras. He innocently points at the zebra and says, “elephant.” In this case, the word “elephant,” if properly understood, means a particular *zebra*. Note that in these cases, I mean by “meaning” and “understanding” the understood usage of the words. Clearly the child is wrong—it’s a zebra not an elephant—the child is errantly using the word “elephant” to refer to the zebra. The pattern that this points out is that, in general, words are understood in *usage* and *usage* entails a certain *context.* Searle (1980), in his description of the Chinese Room, doesn’t talk in detail about the context of symbols, but we can assume the context would be considered in the English rules. The phrase, “squiggle squiggle” may mean something different when in the context of the symbols “lopa mopa” than it does in the context of the symbols “boba laba.” A system that is dynamic enough to use the symbol differently in different contexts more plausibly understands the phrase than one that does not. This doesn’t get us anywhere yet because Searle’s system of symbol manipulation could still provide this ability provided that the English rules apply to long enough sequences of symbols (so that “lopa squiggle squiggle mopa” will lead to different language generation than “boba squiggle squiggle laba” would). Thus, Searle doesn’t hack our intuitions by neglecting context on its own. Rather, he hacks our intuitions by neglecting the impossibility of such a rigid system (i.e. a deterministic one) that still manages to encode all responses to all possible phrases in all possible contexts. This is a preposterous assumption, but one that is necessary for the setup of Searle’s Chinese Room if we are to believe that it could convince a native Chinese speaker that there is another native Chinese speaker in the room.

I would, at this point, like to draw attention to a claim that Searle makes:

> "Notice that the force of the argument is not simply that different machines can have the same input and output while operating on different formal principals—this is not the point at all. Rather, whatever purely formal principles you put into the computer, they will not be sufficient for understanding, since a human will be able to follow the formal principles without understanding anything. No reason whatever has been offered to suppose that such principles are necessary or even contributory, since no reason has been given to suppose that when I understand English I am operating with any formal program at all (p. 4)."

However, this claim that there is no *evidence* that Searle's (or my own) understanding of English is operating with any formal program seems to be wrong. In developmental psychology, there is a well-known phenomenon known as the "poverty of stimulus" which basically says that the speed at which children learn language and the quality of that language is incommensurate with the amount of language they are exposed to (Chomsky, 1986). This seems to imply some sort of innate language capacity or structure. This implication may be disputed, but unless I misunderstand Searle's meaning with "any formal program at all," the proposition that "no reason whatsoever has been offered" is false. Rather, a reason, be it true or false, has been given that such a formal program may exist.

Let me real it back in to the discussion at hand. Context is necessary for understanding. Searle's system accounts for context, but in doing so becomes absolutely infeasible because it would require infinite English rules for infinite Chinese contexts of many (though finite) Chinese symbols. Searle states that no reason has been given for thinking that any sort of formal program might be involved in humans understanding natural language. This is false; the poverty of stimulus is such a reason. Thus, some sort of formal program *could be* involved in human natural language understanding. If such a formal program *is* involved in human natural language, it cannot possibly store infinite rules for infinite contexts. Thus, the program Searle describes is the *wrong one*. It is in this way that Searle hacks our intuitions. He describes the program in such a way that our intuitions can do nothing but deny the possibility of any program by itself being sufficient for intentionality.

Having discussed the power of adding probability to a description of a program and claiming the possibility of some sort of program being involved in human natural language understanding, I now need to develop a description of a plausible program that could be involved in human natural language understanding and, by consequence of being a program, could be involved in machine natural language understanding. I will do so by appealing to three topics in contemporary natural language processing: 1) probabilistic models of language based on context, 2) representation learning, and 3) what I have decided to call "grounded semantics."

Probabilistic models of language are approaches to processing language which use some probability distribution. Some approaches use context to construct a probability distribution (Chater & Manning, 2006). I will describe the intuition and gloss over the details. Imagine you have the text, from earlier "An _____ is an animal with large, floppy ears, a long trunk, and which lives in Africa." Intuitively, you would say that the word "elephant" is *more likely* to fill that blank than, say, the word "helicopter." A probabilistic model will capture this "more likeliness" if the word "elephant" occurs more frequently in the context of words like "animal," "floppy," "ears," "trunk," etc. than does the word "helicopter." Now the question obviously arises—does this constitute *understanding*? Surely it is just pattern recognition. However, the rules aren't hard-coded deterministically. Rather, they are inferred empirically from similar contexts. Thus, at the very least, some sort of empirical inference is occurring. Still, I will, as Searle might, leave it to your intuitions to convince yourself that any program like this doesn't constitute understanding. This system is still symbol manipulation. However, these manipulations are not entirely meaningless. Some of the manipulations mean that word *X* is more likely to occur in context *C* than is word *Y*. This doesn't have anything to do with the *meaning* of word *C*, but the manipulation does have some meaning.

Representation learning is a popular way of thinking about the uber-hot-and-trendy "deep learning." Representation learning is used for all sorts of data, but when applied to language, it is sometimes based on probabilistic language models. It uses artificial neural networks (not to be confused with the neural networks in your brain—artificial neural networks are just mathematical constructs) to develop vector representations of words, phrases, paragraphs, and other pieces of text (Bengio, Courville, & Vincent 2013). The intuition of this is that vectors of similar words, phrases, paragraphs, etc. are near each other (to dab our toes in mathematical lingo, that is to say they have high cosine similarity). For example, we might expect the vectors representing the phrases, "He went to the gym" and "She worked out" to be near each other since they have similar semantics. A famous result of this is meaningful vector arithmetic results. For example, the vector of "King" minus the vector of "Man" plus the vector of "Woman" is very nearly the vector of "Queen" (Mikolov, Yih, & Zweig, 2013). This, at first glance, might seem like we've arrived at understanding. We might buy that this computational system *understands* words and phrases as similar. But, while the system may understand that phrase *A* is similar to phrase *B*, in order to understand phrase *A* itself, it must first understand phrase *B*. In other words, this still does not capture what phrases *themselves* are about, but merely captures *similarities* and *relationships* between phrases. Still, this is a further improvement than before.

Thus, we have two models that are still symbol manipulation (i.e. they manipulate the symbols of natural language): Probabilistic models which model the meaning of words based on their context and representation learning which manipulates symbols meaningfully based on the similarity between one word or phrase and another. These models of symbol manipulation do mean *something*. The problem with each is that the *meaning* is not grounded. We can't really know the meaning of a word given its context unless the meaning of the context is grounded and is not just itself a set of meaningless symbols. Likewise, we can't know the meaning of a phrase given another similar phrase unless one of the phrases is grounded and is not just itself meaningless symbols. As Searle (1980) might say, it still doesn't know "that 'hamburgers' refers to hamburgers" (p. 6).

Thus, the final area of research I wish to discuss is what I'm going to call "semantic grounding[3]." By this, I mean any research that attempts to ground the meaning of natural language symbols in something external. As an example of this, consider the University of Washington's preliminary work in reasoning about size (Bagherinezhad et al., 2016). The research is framed around the question, "Are elephants bigger than butterflies?" The answer is obviously yes. Based on a probabilistic model, perhaps the symbol "elephants" appears before the symbols "bigger then butterflies" with greater frequency than "butterflies" appears before "bigger than elephants." But, as stated before, the symbols still don't have meaning. The interesting thing about this research is that the information of size is externalized in pictures. Thus, if elephants are bigger than butterflies as observed in pictures, then the response to "Are elephants bigger than butterflies?" would be, "Yes." The significance of this is that the answer

[3] The term "semantic grounding" may or may not mean this in the AI literature (I'm not sure, I just heard it in ((Socher, et al., 2013) & (Andreas & Klein, 2014)), but it was a term I thought fit the type of research I am referring to.

wouldn't be determined from just the syntax of the question, even probabilistic syntax, but rather from empirically observed information in pictures of the external world. This research can be thought of as grounding the semantics of language in *visual properties of its referent in a picture*. Thus, a symbol is about the properties of some image. Consequently, it has *some* intentionality. Perhaps this preliminary work could be applied with a higher level of granularity to ground the semantics of "floppy ears" to that of "elephant" instead of "helicopter." The result would be, "floppy ears" refers to some properties of elephants in pictures.

All this discussion of meaningless symbols and grounding semantics opens up the discussion to the philosophical discussion of semantics, which is not the purpose of this essay. Thus, let me make clear what I mean by "meaning." When I say that a program is "meaningless symbol manipulation" I just mean what I think Searle means—namely that there is a distinction between a Chinese symbol to Searle (who doesn't speak Chinese) and the same Chinese symbol to a native speaker. To Searle, the symbol is meaningless; to the native speaker, it has meaning. However, I differ from Searle in this: a symbol can, in various degrees, have meaning to a program in the same way as a Chinese symbol has meaning to a native speaker. In a hard-coded, deterministic system like the one Searle described, the symbol is indeed utterly meaningless to the program. In a probabilistic language model, a symbol has meaning to the program as related to a context (i.e. a symbol is more likely to occur in one given context than another). In representation learning, a symbol has meaning to the program as related to other symbols (i.e. a symbol has similar meaning to another symbol). Lastly, in a grounded system, a symbol has meaning to a program in its reference to an external object (i.e. a symbol refers to some property of an image).

Let me speculate how Searle would respond. He says that there are three different question. First, “Could a machine think?” which he answers, “Yes. We are precisely such machines” (1980, p. 11). Second, “Could a digital computer think? If by ‘digital computer’ we mean anything at all that has a level of description where it can correctly be described as the instantiation of a computer program…” (p. 11). Which he again answers, “Yes, since we are the instantiations of any number of computer programs, and we think” (p. 11). Third, “But could something think, understand, and so on solely on virtue of being a computer with the right sort of program? Could instantiating a program, the right program of course, by itself be a sufficient condition of understanding?” which he answers, “No, because the formal symbol manipulations by themselves don’t have any intentionality; they are quite meaningless” (p. 11). He also states that the third question is often confused with one of the first two. The distinction, according to Searle, is that the second question pertains to some system which can be considered as a program, but which has some understanding due to the fact that it does something more than symbol manipulation, which is assumed meaningless. On the other hand, the third question pertains to a system which is also considered a program, but which does nothing more than meaningless symbol manipulation and consequently doesn’t constitute intentionality. I think he would respond to my argument that I am confusing the third question with the second. The level of description that I have given, which is a fairly high level one, can be described as the instantiation of a computer program, thus Searle might argue that I am actually just presenting a description of a program to answer the second question while the response to the third question remains, “no.” I respond to this objection with the claim that the third question *is* the second question; hence, the “confusion” would actually be to think of them as different questions. Why are they the same question? Because the symbol manipulations, contrary to Searle’s assumption, *do* have meaning. The manipulations relate words to contexts, words and phrases to other words and phrases, and words to external objects. They don't have *much* meaning yet. But this is more than no meaning. And if symbol manipulation can have meaning, then the third question becomes the second and even Searle agrees the answer to that question is yes! That is, if symbol manipulation can have meaning, then it is coherently possible that instantiating a computer program can be sufficient for intentionality. I have described a system of meaningful symbol manipulation. Thus, it is coherently possible that instantiating a computer program can be sufficient for intentionality.

References

Andreas, J., & Klein, D. (2014). Grounding Language with Points and Paths in Continuous Spaces. *Proceedings of the Eighteenth Conference on Computational Language Learning* (pp. 58–67). Baltimore, Maryland: Association for Computational Linguistics.

Bagherinezhad, H., Hajishirzi, H., Choi, Y., & Farhadi, A. (2016). Are Elephants Bigger than Butterflies? Reasoning about Sizes of Objects. *Association for the Advancement of Artificial Intelligence*.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1798-1828.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 335-344.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use* (Convergence (New York, N.Y.)). New York: Praeger.

Chomsky, N. [Stony Brook University]. (2009, May 21). *Noam Chomsky: The Stony Brook Interviews Part Two* [Video File]. Retrieved from https://www.youtube.com/watch?v=CHS1NraVsAc&t=95s

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 746–751). Atlanta, Georgia: Association for Computational Linguistics.

Putnam, H. (2002). The Nature of Mental States. In D. J. Chalmers, *Philosophy of Mind: Classical and Contemporary Readings* (pp. 73-79). New York, New York: Oxford University Press.

Rescorla, M. (2015, October 16). *The Computational Theory of Mind*. Retrieved March 19, 2018, from https://plato.stanford.edu/archives/spr2017/entries/computational-mind/

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Oxford, England: Lawrence Erlbaum.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 417-457.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., & Ng, A. Y. (2013). Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Transactions of the Association for Computational Linguistics.* Association for Computational Linguistics.