Aporia vol. 19 no. 2–2009

The Chinese Room: Qualia and Semantics

Joshua Mitchell

Throughout the literature generated by Searle's Chinese Room Argument, the Robot Reply has persistently beckoned a more thorough response. Many proponents of AI have stood their ground in holding that there is something intuitively correct about such a reply.[1] For example, critics such as Bridgeman and Abelson contend that if a super robot were able to interact with the world, then this should generate understanding, as is seen in children. (Abelson actually uses the term "sensorimotor" for the phenomena needed for the machine to do such a thing (424).)[2] It is this requirement that I shall address directly. *Assuming* that the utilization of robotics is encompassed by the school of strong artificial intelligence, I will contend that this can take the system in question only so far. Indeed, the robot will never understand any concepts that directly involve qualia, despite any newfound sense of sensorimotor phenomena.[3] In fact, we will see that even if the strong AI school is able to account for semantics

[1] By A.I., throughout the entirety of this paper I am referring to strong AI; i.e., the idea that a physical symbol system is both necessary and sufficient for intelligence.

[2] By sensorimotor, I am referring to phenomena such as sight and hearing that allow us to associate a concept's given syntax with its content via these sensorimotor skills (e.g., this furthers one's understanding of an apple if she can mentally picture an apple because she has seen one previously. This remains true for other sensorimotor phenomena such as sound, as it heightens one's understanding of terms such as "melodic minor scale.")

[3] For an explanation of qualia, see section II of this paper.

Joshua Mitchell is a senior from the University of Virginia majoring in philosophy as well as Middle Eastern languages and literatures. His interests include Arabic and Persian poetry, philosophy of mind, ethics, and philosophy of religion. Upon graduation, he will be commissioned as a Naval Officer in the United States Navy.

using only syntax, this inability to understand qualia will prevent a robot from fully understanding many terms that inadvertently involve qualia.

In what follows, I will describe the Robot Reply in response to the Chinese Room Argument and argue that qualia are necessary to fully understand terms that refer to the qualia themselves. Furthermore, I will argue that even if strong AI were able to achieve semantics solely on syntax, the symbol system's inability to observe qualia would still preclude human-level understanding for the relevant terms associated with the respective qualia. Finally, I will explore how many terms and concepts this truly affects.

I. The Whimsical Robot Reply (and What it is Missing)

First, it is crucial to note that the Robot Reply, even taken at best, is not within the grasp of strong AI in its original formulation. The doctrine of strong AI explicitly states that "the appropriately programmed computer really is a mind in the sense that computers given the right programs can literally be said to *understand* and have other cognitive states" (Searle 417–18). This suggests that all that constitutes understanding is what can be done by syntax alone. However, for the sake of argument, let us *assume* that strong AI can appeal to sensorimotor phenomena. What, then, would this offer to Newell and Simon's successors?[4]

As Bridgeman points out, "the robot can internalize meaning only if it can receive information . . . with a known relationship to the outside world" (428). In other words, if the robot has sensorimotor connections with the external world, then the correlations between the entity being understood and the external information can be realized, which will inevitably aid in some type of understanding (in particular, external information that allows for internal pictorial representations). Although Searle ultimately contends that this is not true *understanding*, Bridgeman attempts to demonstrate that it is.[5] In doing so, he makes a fair analogy of a robot learning the number five as would a child, "where the occurrence of the string of symbols representing 'five' in its visual and auditory inputs corresponds with the more direct experience of five" (427). Thus, if this type of association is what enables children to learn, then this becomes plausible for the robot. Consequently, it is implicative of different tiers of understanding. Surely, this would correspond with the different tiers of intelligence observed in animals.

[4] Newell and Simon are often considered the main architects of strong A.I. For a comprehensive description of their thesis, see Haugeland *Mind Design II*.

[5] We assume, of course, that Searle's concept of understanding is a human one.

However, lest we forget, the doctrine of strong AI demands that the system's (in this case, the robot's) understanding must be comparable to that of the human mind.[6] Thus, if there is some aspect of human-level understanding that the system will never achieve, then strong AI has ultimately failed. And while the allowance of robotics in the classification of strong AI seems to mirror *some* understanding comparable to humans, *prima facie*, there is in fact a phenomenon that is crucial to the human level of understanding that the robot will almost certainly never have: the phenomenon of *qualia*.

II. Qualia

Undoubtedly, one may be skeptical of the idea that qualia are needed to understand concepts that refer to the individual quale (e.g., to fully understand what the general term "fear" is, one needs to have had a *fear* quale). However, let us consider that qualia by definition are the raw feels of experience (i.e., the "what it is like" of experience) (Nagel 1974). Under Nagel's scrutiny, we find that these "subjective characters of experience" are defined and understood solely by our experience of them (which is why they are unique to species, and arguably person specific as well. Thus, if one has never experienced a certain quale, then it is difficult to believe that one truly understands the meaning of any terms directly related to that quale.[7] For example, suppose Billy is a small child who has been raised in a utopian society where there exists nothing that invokes fear. Now, Billy continually reads about *what it is like* to be afraid, so he knows someone who is afraid becomes light headed, clammy, slightly nauseous, etc. Perhaps we can even venture to say that he *understands what fear is to a certain extent*. However, if Billy is released into the real world and someone jumps out in front of him dressed as a monster, he certainly understands *fear* more now than he did before.[8]

There is another way to argue that qualia contribute to the meaning of their relevant qualia terms. We can picture some world in which there

[6] Newell and Simon claim that being a physical symbol system is both a necessary and sufficient condition for exhibiting intelligence, and intelligence is taken to be in the domain of the human.

[7] Note that when I say one does not truly understand a term's meaning, I mean that they do not understand it at the optimum level, i.e., there is another higher and reasonably obtainable tier of understanding still left to reach.

[8] There are certain parallels between this and Jackson's thought experiment on Mary's Room (291–95). However, I am not taking a stance between physicalism and dualism here. What is important to see is that the presence of a quale has heightened, to some extent, one's understanding of a given concept—likewise, we could always conjecture that when Mary leaves the room she does not learn new facts, but rather, her understanding of certain facts is heightened.

is a set of physical duplicates that have inverted qualia.[9] It seems dubious at best to assert that if the duplicates have the identical belief "I am overjoyed," but one is really experiencing the melancholy quale, that they really mean the same thing. Therefore, since the meaning of the identical sentence has been altered by the presence of a particular quale, we can see that qualia must have *some* influence over the meanings of their respective qualia terms.

Thus, we may say that to *fully* understand things involving any given qualia term (e.g., being red, scary, joyful, lustful, the note C#) requires that one experience the relevant quale (e.g., hearing the note C#). We can even see this as true for more complex concepts. Take for example *claustrophobia* and *altruism*. Claustrophobia is, of course, fear of enclosed spaces. In order to understand the word "claustrophobia," one would have to understand each constituent term which that word encompasses (fear, enclosed, spaces, etc.). However, since fear is a quale, one must experience fear in order to fully understand the term. The same reasoning goes for "altruism" and other seemingly complex words.

Thus, so far we may state the following:

> 1. To fully understand concept C requires understanding all sub-conceptual constituents of C (namely *C1–CN*).[10]
>
> 2. If any of the constituents [*C1–CN*] are in fact a qualia *QP*, then one must also fully understand *QP* to fully understand the set [*C1–CN*], and consequently C.
>
> 3. The only way to fully understand *QP* is to experience *QP*.

Still, a proponent of strong AI may contend that if a robot is capable of sensorimotor phenomena, perhaps the robot can *realize* its own set of qualia by experiencing the world through these phenomena. While this initially seems to be a reasonable line of thought, there are some difficulties that unavoidably meet such a claim. For this to be feasible, qualia must not be imbedded biologically (i.e., they must not be exclusive to biological beings). However, there is some empirical evidence that seems contrary to this. Various medical observations indicate that people who have a certain

[9] This is, of course, assuming this is possible for the sake of argument. I am in no way taking a stance on whether or not this is actually possible, but merely wish to demonstrate the influence of qualia on meaning.

[10] The "–" meaning "through," not "minus."

genetic trait do not feel the *pain* quale.[11] This may suggest that qualia are deeply imbedded in some biological phenomena. Of course, this does not mean more research is not needed. It merely shows that there are in fact correlations between biological factors and certain qualia. However, aside from looking to scientific data, there is a much more general concern. We have absolutely no idea *how* the human brain produces qualia (we certainly have ideas, but nothing is for certain), let alone how to reproduce them in a machine. Thus, it appears the burden of proof is on those who claim that machines could have qualia. This is not to say that one day this will not occur. Inarguably, more empirical research will answer this, as this specific issue is an empirical one. However, surely one would concede that as of now it is a notion which is quite dubitable.

Thus, let us add to our summary:

> 4. Assume that all qualia Q are biologically embedded.
>
> 5. Machines are not biological, thus cannot have biologically embedded entities.
>
> 6. Therefore, any C which includes any qualia Q as a constituent cannot be fully understood by machines.

One will certainly note that there is a very large assumption that plays a pivotal role in the argument thus far (i.e., that all qualia are biologically embedded). I will acknowledge it as such, and thus perhaps my argument is somewhat modest in nature because of it. However, one will also note the above justification for such an assumption at this time.

III. The Importance of Qualia from Another Perspective

To further my point on the importance of qualia in fully understanding concepts, I am going to make an allowance to the proponents of strong AI. Let us *assume* that perhaps one day the strong AI programmers successfully generate significant semantics solely from syntax. In this case we can imagine their reasoning as follows: Perhaps there are a handful of qualia terms (and even some words or concepts such as claustrophobia) that the system will not be able to understand at the human level. However, not only are there words which the robot can understand that only depend on sensorimotor phenomena for understanding (e.g., simple associations between a syntactic expression and phenomena such as

[11] Specifically, the lack of the SCN9A gene. See Woods.

inner-pictorial representation), but there are also many words where the need for such phenomena is simply implausible. Therefore, the "deficiency" of attaining human level understanding of language is a small one.[12]

Yet, one must be careful when making general claims such as these. Indeed, after much scrutinizing, what initially seems quite certain is rather quite the opposite. For example, to many the concept of *neutral evolution* seems to be merely a fact—just a definition of a relatively abstract fact. *Prima facie*, this requires neither inner sensorimotor phenomena nor qualia. However, I will show that this example collapses under examination, thus weakening the objection.

Let *T* be any given complex or abstract term (like neutral evolution).[13] More likely than not, *T* can be broken down into simpler terms or concepts *W*, *X*, *Y*, *Z*, etc., of which *T* consists (e.g., neutral evolution can be broken down into concepts of species, mutation, fitness, advantage, and so on). These terms *W*–*Z* . . . will either be irreducible, or they will lend themselves to being broken down into simpler terms of which they consist. Eventually, all the terms or concepts will be at an irreducible level. So long as *at least* one of these terms or concepts requires either qualia or sensorimotor phenomena in order to be understood (and one would assume the number would be much greater than one term), *T* cannot be understood *fully* without such phenomena.[14]

In our example, the abstract term "species" can be broken down into mammals, reptiles, and so forth, and of those I have a mental image of a stereotypical animal.[15] Arguably, if I do not possess the ability to have an inner pictorial image of an animal when I think *species*, I cannot truly understand what a species is.[16] If so, I will not have a full understanding of what neutral evolution is. We have already granted that perhaps the robot

[12] I would like to thank Professor Paul Humphreys for raising this objection in one of our many discussions as well as the example of neutral evolution.

[13] Note that I am not considering things that are holistically mathematical in nature. I am only considering terms and concepts which in general are semantically reducible, given the presented procedure above.

[14] It is necessary to point out once more that our goal is to understand a given term or concept to the highest degree reasonably possible for a human, as the human is the standard for the mind being emulated by a machine. If the machine cannot do this, then strong AI has arguably failed.

[15] See Minsky.

[16] Such an idea can be gathered from the externalist point of view. If I never see a "mother," then I cannot have an inner image of "mother" in my head when I think of her. If this is true, then surely I do not have a *full* understanding of what "mother" entails. If nothing else, we would say that one who can picture a "mother" in their thoughts has a fuller understanding of what "mother" is then one who does not possess such an image.

does have these inner-pictorial representations, so for the sake of argument we may conjecture that the robot can understand thus far, in that it is able to correlate the term "animal" with some image. However, we have previously established that qualia will not be present. Thus, when the robot is met with a sub-term requiring qualia, it will run into some trouble, as is the case in this example. Considering sub-terms pertaining to neutral evolution, "advantage" is certainly one of them. The concept *advantage* entails the concept of *success*, which in turn entails some raw feeling of achievement, even if only at a subconscious level. If one cannot *fully* understand this raw feeling of achievement, then one cannot fully understand what success truly is. This would certainly be the case if one was not aware of this quale, as one recalls from the previous example of Billy and fear. If this is the case, then one does not have the optimal understanding of *advantage*, and so on. Some cases are more obvious than others, but seemingly there is a difference in the level of understanding between an agent who has not experienced the relevant quale (or qualia) and an agent who has. And since we are dealing with the mind, per the doctrine of strong AI, we must use the average human being as our standard of evaluating a system's satisfactory level of understanding. Regardless of how small the difference is, so long as the average human can understand it, the machine *must* follow suit.

IV. The Objection of Partial Understanding

Perhaps one might object concerning the very fact that we are using the "average human" as the standard of understanding. Indeed, the AI proponent might claim that even humans can understand certain concepts without having to *fully* understand every constituent sub-concept therein. For example, when a high school student takes an introductory calculus course, surely he does not understand *every* concept of integral calculus, but we would certainly want to claim that he *understood it* to a satisfactory, *average* degree. If this is so, then can the robot not reach an *average* understanding which bypasses some of the more difficult and complex sub-concepts of a given topic?

Such an objection may seem reasonable, but here there is a minor discrepancy. First, let us imagine a concept C with sub-conceptual constituents forming a hierarchical pyramid. The most crucial sub-concepts are at the bottom, whereas the more detailed and specialized concepts are at the top. Obviously, the foundational concepts are the most important to have in order to achieve any kind of understanding. Now, in the calculus objection, the student may not know *every* concept and still have an average understanding of the subject, but certainly she has almost all (if not all) of

the foundational sub-concepts. Concomitantly, we will assume that even if we said that it were *possible* for a robot to achieve average understanding with only *some* of the sub-concepts being understood, surely we would all concede that they would need to be the foundational sub-concepts. However, how many concepts have qualia as part of their "fundamental" sub-concepts?

Anything that pertains to romance or love certainly has qualia terms as fundamental concepts (such as affection, happiness, sadness, etc.). Anything that has to do with seeing colors, or hearing things will also follow suit. Furthermore, we note the breadth of inclusivity here. For example, with colors this extends from understanding what we mean in talking about rainbows, to describing optics and wavelengths in physics. For hearing, this extends from simply discerning the tunes a bird sings, to understanding the meaning of beats generated by sound wave interferences. We notice a pattern that things involving sensory phenomena find themselves easily on this list, and in fact, there is much in our world which invokes terms pertaining to sensory phenomena. These things seemingly require an understanding of their respective qualia upfront in order to have a fundamental comprehension. I leave it up to the reader to think of other explicit examples, but feel that we may agree that the number we could eventually generate would be large enough to drastically reduce the force of this objection.

V. Conclusion

In light of this discussion, what should one take away from such discourse? I will highlight what has been presented in this paper below.

First, the Robot Reply, although contrary to the original doctrine of strong AI, only allows the system to have *dumtaxat* minimal understanding of words or concepts where only associations between syntax and internal representations through sensorimotor phenomena are needed. However, it fails to account for the qualia needed for a full, *human-level* understanding of the world. Second, even if Searle is completely wrong in his Chinese Room Argument, and we postulate that one day computer programming could allow for semantics to be derived from syntax, qualia would still not be present—thus posing the same barrier as in (1). Third, as of now, the burden of proof with computers generating qualia is with the programmers, granted that it is not even known how qualia emerge in the human species. And fourth, the problems that arise from the first and second point thus seem to be inescapable.

Because language is the model that strong AI uses for cognition, it must also be its barrier. If there are areas of language that cannot be accounted for, then arguably there will be areas of cognition that are unaccountable in the strong AI model as well. In particular, it seems that any part of language that *at some level* (and more often than not at some *foundational* level) could require semantic understanding of qualia is out of strong AI's grasp. This is because the algorithms used in computational systems are linear. Therefore, if there are issues that are unaccounted for at the base of a concept or combination of syntax, the end result (which in this case in understanding) will inevitably fail to achieve a human level of understanding. Thus, the question we must now pose is *how much language will be inhibited?*

As I have pointed out, any words or concepts that have to do with colors, emotions, or feelings require qualia in order to achieve their *full understanding* (so these are "off limits" to AI). However, any word whose sub-constituent words or concepts involve qualia will also be forbidden in terms of an AI system's full understanding. This includes the more obvious terms such as "claustrophobia," but also may include more counterintuitive concepts such as neutral evolution. When all these words and concepts are tallied up, I believe they will be a significant amount.[17] Perhaps this is an empirical question—and if it is, then we must conduct further linguistic research in order to determine how strong a limit AI is working against.

Computers, in their history of "evolution" within the past fifty years, have irrefutably attained many respectable and awe-inspiring things. However, it has been the purpose of this paper to demonstrate that one should always be skeptical of the notion that a non-living machine could ever possess qualities that are (so far as science has told us) exclusively intrinsic to higher *living* and *breathing* organisms. My position can thus be seen as a medial one between the one extremity of Searle's thesis (being that computers will never be able to understand any semantic content) and the doctrine of strong AI, which subscribes to the exact opposite view. While computers may at one point be able to have some elementary level of *understanding*, so long as qualia are deemed to have some significance within the realm of understanding (which, I believe they always will), computers will ever so ceaselessly remain, as Hubert Dreyfus so elegantly puts, "at an impasse" (143).

I would like to thank Jason Megill for all of his suggestions and critiques of this paper.

[17] Although perhaps not in the sheer number of unique terms, but in the frequency that they either appear in sub-constituent concepts and as "qualia terms" themselves.

Works Cited

Abelson, Robert. P. "Searle's argument is just a set of Chinese symbols." *Behavioral and Brain Sciences* 3.3 (1980): 424

Bridgeman, Bruce. "Brains + programs = minds." *Behavioral and Brain Sciences* 3.3 (1980): 427.

Dreyfus, Hubert L. "From Micro-Worlds to Knowledge Representation: AI at an Impasse." *Mind Design II*. Ed. John Haugeland. Cambridge: MIT P, 1997. 143–82.

Jackson, F. "What Mary Didn't Know." *Journal of Philosophy* 83 (1986): 291–95.

Minsky, Marvin. "A Framework for Representing Knowledge." *Mind Design II*. Ed. John Haugeland. Cambridge: MIT P, 1997. 111–142

Nagel, Thomas. "What is it Like to Be a Bat?" *Philosophical Review* 83 (1974): 435–50.

Newell, Allan, and Herbert A. Simon. "Computer Science as Empirical Inquiry: Symbols and Search." *Mind Design II*. Ed. John Haugeland. Cambridge: MIT P, 1997. 81–110.

Searle, John. R. *Minds, brains, and programs*. Behavioral and Brain Sciences 3.3 (1980): 417–24.

Woods, Geoff. *Cambridge Institute for Medical Research* as reported by the Cambridge University press release on December 11th, 2006 <http://www.admin.cam.ac.uk/news/press/dpp/2006121102>.