The Extended Room
or
What Otto Didn't Know

Ryan Victor

1. INTRODUCTION

The Chinese room example is one of the most famous and important counterarguments to the 'strong AI' thesis, based on Turing's work, that any computer capable of human-like dialogue understands English and has thoughts much like our own. The systems reply has been levelled against the Chinese room example as a way of defending this claim. I propose that the systems reply fails because the heart of its objection is based on a non-standard definition of "understanding", and does not consider the concept of consciousness. Furthermore, I will argue against the Extended Mind thesis, as advocated by Clark and Chalmers, in which a redefinition of the boundaries of the mind is proposed. I will claim that the extended mind thesis is very much like the systems reply. It fails because it creates entirely new, non-standard concepts of "mind" and "belief" that are used equivocally with the standard definitions, does not factor in consciousness, and the criteria proposed for being part of the mind are no less arbitrary than those that they are intended to replace.

Ryan Victor, originally from London, is an undergraduate Philosophy major at the University of Pennsylvania. His interests within Philosophy include Philosophy of Mind, Philosophy of Science and Philosophy of Law. Ryan stresses interdisciplinary study and has written in areas such as Psychology and Evolutionary Biology. He will soon be publishing in the field of Ethnomusicology, focusing on the historical trajectory of Jamaican music and culture, and its interaction with that of Britain during the latter half of the 20th century. Ryan is also an avid composer of electronic music. After graduating from Penn, Ryan plans to attain a degree in Law and pursue his lifelong ambition of working to improve social justice, equality, and international human rights.

2. THE CHINESE ROOM AND THE SYSTEMS REPLY

Proponents of 'strong AI' make the claim, based on Turing's groundbreaking arguments in *Computing Machinery and Intelligence*,[1] that any computer capable of fooling a human into thinking that they are conversing with another human in a written (or on screen) dialogue would have thoughts and understanding in the same way as we do. Thus, any understanding of the nature of the computer program that achieves this gives us understanding of the workings of the human mind. Furthermore, the argument has been made (e.g. by Block[2]) that the necessary requirement for thought, which both we, as humans, and the computer that can pass the Turing test share, is the ability of formal symbol manipulation. In *Mind, Brains and Programs,* Searle[3] creates a thought experiment in an attempt to show (contra Shanck and Abelson[4]) that formal symbol manipulation is not a sufficient condition for thought, and thus passing the Turing test is not a guarantee of a computer's understanding of language.

The thought experiment is very simple. Searle is in a room and is passed cards with Chinese characters on them. He has a look up table in a giant book, which tells him what to write as a response to any given set of characters. He passes his answer back, and the person on the other side, who only sees the replies written on cards, believes he is having a legitimate conversation with whomever is inside the room. Since Searle does not speak Chinese, it is absurd to claim that, armed with his rulebook, he suddenly understands the contents of the discussion he is a part of. This is evidenced by the fact that, were the interlocutor to pass through a message telling Searle that the room is on fire, although he might pass a card back which says "thanks very much, I'll leave," Searle would not actually do so – how could he?

This experiment, then, can be thought of as analogous to a computer, programmed with a look up table of English sentences and rules for applying them. In both cases, there is indeed formal symbol manipulation, and in both cases it is certainly possible that a human may well be fooled into believing that they are part

of an interesting conversation with a conspecific, but the fact remains that, in both cases, there is no *understanding*. Searle himself is acting in exactly the same way as the CPU – simply following instructions blindly. We may think of another analogy. If a heat-seeking missile is programmed to lock on to a plane, a naïve observer might describe the missile's behaviour as though it "wants" to hit the target, but this talk of the missile as an intentional system is incorrect – the missile only has *derived intentionality*. In the same way, one might say that Searle in the Chinese room only has *derived understanding*, since the only thing that actually adds to the understanding of Chinese in the room is the rulebook, which was written by someone who actually does speak the language.

The systems reply is an objection to Searle's thought experiment in which the term 'understanding' is applied not to just Searle, but to the system of the room, the book, and Searle as a whole. The counterargument simply states that it is not the case that Searle understands Chinese, but that the system as a whole does. As Block puts it, "If the whole system understands Chinese, that should not lead us to expect the CPU to understand Chinese." Searle's response is a good one. Imagine that Searle simply internalises the rule- book by memorising the entire thing. Remove the room. Let the entire process be in Searle's head. Searle still does not understand Chinese.

The major error in the systems reply is that the word 'understanding' is being used to mean something completely different to how we normally define it, and its redefinition is not made explicit at the outset. Rather, the standard definition along with the extended definition are used interchangeably. There are many systems composed of multiple parts that function as a whole in modern society. For example, the postal system is made up of numerous complex pieces. Whereas we might say that the postman understood where to deliver a letter, would we say, "the Postal System understood that my letter needed to get to Kentucky in two days"? Absolutely not – we do no speak of systems in this way. Understanding is limited, in everyday usage, to, at the largest, the consciousness of an individual organism.

Understanding, were it to occur in machines, would have to occur at the level of the machine's motherboard, including its memory and CPU. To claim that a word like "under standing" can be extended to objects that are connected in such a cursory manner is contrary to both intuition and common usage.

It might be countered that we do speak in this way occasionally, for example, "the FBI knew the man way headed to Venezuela" or "NASA announced that the launch will be postponed." Although we do accept these kinds of sentences, they are a matter of common parlance and not a matter of conceptualising the system as a whole. We could easily pinpoint the individual agents within the FBI who knew the fugitive's whereabouts, and similarly a NASA statement is written by individual people who have an understanding of the NASA mission as a whole. These individuals would still have the same understanding absent the agencies for which they work, whereas the room, book and Searle system requires both the rulebook and Searle together, and even then they add up to create nothing more than an *illusion* of understanding.

A second criticism is that the systems reply takes no account of consciousness. Searle and the rulebook cannot be thought of in terms of a system because the rulebook is not conscious, not does it play an active or direct role in Searle's consciousness. Thus, they cannot be thought of as a single system. The rulebook is acting as a store of information with which Searle can interact in order to produce the correct responses, but there is an extra layer of translation and manipulation that only Searle, and not the rulebook alone, can accomplish.

3. THE EXTENDED MIND

Clark and Chalmers' thesis of "the Extended Mind"[5] involves the question of exactly where the mind ends, and where to draw the boundaries. This question is addressed by recourse to a couple of examples. The first involves a Tetris-like set of scenarios. Let us imagine a situation where, in order to ascertain whether a certain shape block will fit within a group of differ-

ently shaped blocks (I assume we've all played Tetris), we can either mentally imagine it, or use a button (which can do it faster) to rotate it. Clearly, we imagine the first case as a mental rotation, and in the second the rotation is clearly external. Now, imagine a case in which the ability to perform the fast rotation is implanted into our visual system, or, if that is too far a stretch, that the rotation can be activated on screen by thinking about it. (This is technology currently under development to assist the disabled.[6]) This seems to be a bit of both. To try and solve the problem, Clark & Chalmers propose the 'parity principle', which states that, "If, as we confront some task, a part of the world functions as a process which, were it to go on in the head, we would have no hesitation in accepting as part of the cognitive process, then that part of the world is (for that time) part of the cognitive process."[7]

The second example involves a man, Otto, who has memory loss, and thus must write everything in a notebook in order to be able to function normally. Thus, in a case where he is asked what he believes about a certain topic (e.g. where is MoMA) he must consult his book before being able to answer the question. Clark et al claim that this means the notebook is as much a store of his beliefs as the internal part of the brain in which that information might otherwise be stored (the hippocampus?). Thus, we should treat Otto storing his belief about MoMA in his book and me storing my belief in my brain as exactly the same. The conclusion of these two examples, then, is that any physical system that is recruited as part of a cognitive process is, by virtue of its participation in that process, *part of the mind itself*. Furthermore, an agent's beliefs, beings as they are part of cognitive processes, can be stored in any physical location, outside or inside the brain. The criteria supplied for judging whether or not something like a notebook would be a candidate for a part of Otto's mind are that it is consulted often, the contents are always or for the most part instantly affirmed, and the resource is easily consulted.

In support of this argument, an analogy can be drawn between adding extra memory to a computer and the notebook example. An external hard drive should not be viewed as any more

or less part of the computer than an internal one, so why should we favour our own internal grey matter over an external memory storage device such as a notebook? A bias towards seeing anything external as automatically not part of the mind is not an *a priori* reason to discount the view. Further, one might suggest that an actual part of the brain could be transplanted outside the body, kept in a vat, and connected via radio contact with the brain. In this case, the external item is performing cognitive operations, and so what, apart from the fact that it is outside the actual skull, separates this from being part of the mind? In this case, the answer is nothing.

4. OBJECTIONS AND COUNTEREXAMPLES TO THE EXTENDED MIND

To try to drive a wedge between the two examples above, let us consider a further example – a watch. During the day, I may consult my watch multiple times per hour, I will instantly affirm that the time represented on it is correct, and it is incredibly easy to consult – a turn of the wrist is all that is required. Thus, a watch satisfies all three of Clark and Chalmers' criteria. Would we consider the watch to be part of my mind? If an adherent to the externalist view would still claim that it is, then let us push the example further. Imagine a time before wristwatches. People in earlier times consulted the sun in order to figure out the time of day. They could do this multiple times per day, there would be no reason to doubt the position of the sun, and what is easier to do than to simply look up. Thus, consulting the sun for the time seems to satisfy the exact same criteria as using a wristwatch. By the same logic, then, it seems that the sun, too, is part of our mind. Even a dogmatic believer in externalism might want to reconsider this assertion. It seems as though the criteria can easily lead us too far astray.

Clark et al's assertion is that our intuitions are deceiving us into using an arbitrary boundary, the skull, in order to define where the mind ends. It is perfectly reasonable to insist that this distinction is not as easy as it seems to be, and questions of exactly where we draw the line are valid. However, in place of this

distinction, they substitute another equally arbitrary criterion – a high degree of reliance, or "strong coupling". The same question of where we draw the line is equally, if not even more, relevant in this case. After all, what constitutes as exactly enough reliance – consulting the notebook once an hour, once a day, once a week?

A further and deeper point is that our intuition seems to strongly disallow the notebook, but allow the external brain segment. Why? I propose a different concept of "strong coupling". Rather than meaning degree of reliance, I would use "strong coupling" to describe the degree to which the actual cognitive process is ingrained and connected to the external device. The reason why the notebook seems so implausible is twofold – firstly, it requires the beliefs to be translated into natural language, and then into symbols by the person before being recorded, and then the inverse when the information is accessed. This need for translation has a decoupling effect – the integration into the cognitive process seems less and less direct as we add further stages of translation and manipulation. Secondly, a notebook is something that anyone can have access to – this is something we would not claim for the information stored in our hippocampus. Something that were connected in a more direct way seems more plausible as a candidate for a part of the mind.

The watch and the sun example seem implausible because they are being considered as part of the cognitive process in the same way as the cerebral cortex may be considered part of the cognitive process, but these are two very different things. Checking one's watch, or the angle subtended by the sun involves consulting aspects of the external world in order to aid in belief and desire formation. For example, knowing that the sun is directly overhead can cause the beliefs that it is around noon, and thus the desire to begin making lunch. However, the actual cognitive processing itself consists in the interpretation of the sun's angle as a functional isomorphism for a particular hour. Like the Chinese room rulebook, the sun itself is not actually performing that interpretive task, and thus it is not part of the mind. The distinction here is between the attributes of the physical world that, as

humans, we are capable of interpreting as representing important information needed to form beliefs and desires, and the actual substantive matter that provides us with the ability to create and manipulate these beliefs and desires in order to produce action.

To try to provide a counterexample that more closely parallels the general logic of the extended mind argument, let us imagine exactly the inverse scenario. I wish to try and define exactly what is means to be 'part of a computer'. Normally, we would say that the computer ends where the boundaries of the machine are, yet this may simply be a bias we have. Surely, if I were experiencing a difficulty with a program in that it took up too much space on the hard drive, I could conceivably print out the information, and then, when necessary, input the code back into the machine when the program required it. This would mean that the computer's memory is not only located inside the machine, but also on the page. Why should we say, then, that the paper is not part of the computer? After all, it requires the information on the page every single time it runs a specific program. We should not bias this kind of storage space simply because it is external, should we? Now, what if I simply memorised the code? Then, wouldn't I be part of the computer too? Again, our intuitions scream no, and for good reason. Neither the actual piece of paper nor my mind are coupled, or connected, in the same way as the actual circuits in the computer are. There is an extra decoupling stage, the printout, or the memorisation of the code, which should automatically disqualify those things from being part of the computer. In the same way, this kind of decoupling should automatically disqualify a notebook as a candidate for part of a mind.

5. ANALYSIS

The reason Clark & Chalmers seem to be able to argue with some conviction and success is because the very term "mind" is being used in two distinct ways. They are guilty, then, of the fallacy of equivocation. The first kind of "mind" is the kind we all understand when we say the word – my mind is the seat of my

consciousness, contained within my brain, or the parts of which work together to perform all of the necessary cognitive processing, of language, perception, memory, etc. that I require to function. The second term, "mind" is used to denote the total of my knowledge, experience and beliefs. This is not the same as the first kind, because much of my knowledge exists in other places; my experiences have been shared by others, and my beliefs, by virtue of being expressed in language, can exist separate from me (as represented on a page, for example). But to equate these two things – the actual hardware necessary to be able to have beliefs in the first place, and the beliefs themselves (articulated or otherwise), is to equate apples and oranges.

A further point is that the notion of 'belief' is also being confounded. In one sense, I would call my mind the store of my beliefs, but in a deep way, since it is the hardware that enables me to have beliefs in the first place and manipulate them in a holistic manner. An essay that expresses my opinion on some matter is also a store of my beliefs, but in a very different sense. In the case of the beliefs in my head, these may be very inarticulate and vague – many people have beliefs that are logically inconsistent. But when we actually have to pronounce on a topic, we formulate our beliefs into natural language, and these become codified into assertions. In many cases, the pressure of having to pronounce a particular belief may cause us to say something we later regret.[8] Thus, the beliefs on the page and the beliefs in the head are not one and the same. Secondly, we often speak of *degrees* of belief, a concept for which we seem to have an intuitive grasp when talking about beliefs in our head, but for which there is no easy expression in natural language. The example of the location of a museum blurs these distinctions because it is too simplistic – many beliefs are far more complex. A further point is that expressing beliefs In a notebook prevents them from interacting in the holistic way that beliefs in the brain do? How can they when they are fixed representations on a page. Thus, it is not as easy to claim that the notebook is as legitimate a store of beliefs as the mind.

The final point is, like the systems reply, the extended

mind fails to take account of consciousness as a legitimate criterion for judging whether something is part of the mind. A common definition of the mind is the 'seat of consciousness'. Thus, anything that can be legitimately considered part of the mind must either contribute to, or be necessary for, this consciousness. That is why we can accept that the brain segment in the vat is still part of the mind, despite being external to the skull, whereas the notebook is not, because the notebook contains no elements that we would call conscious, or in any way can contribute to any part of the intrinsic consciousness of Otto, or anyone else.

6. CONCLUSION AND FINAL THOUGHTS

After outlining the Chinese room, the systems reply and the extended mind, parallels were drawn between the systems reply and the extended mind thesis in order to articulate the drawbacks of the extended mind approach. These included, in particular, the failure to adequately factor in consciousness, and the fallacy of equivocation in relation to the definition of "understanding" and "mind". With the use of counterexamples, the inadequacies of the 'strong coupling' criteria were presented, concluding with the judgement that they are as arbitrary, if not more so, then the criteria they are intended to replace. Finally, alternate criteria, such as a version of 'strong coupling' based on the connection and integration into the cognitive process were proposed.

In general, the extended mind may first appear as a radical and exciting new argument for externalism. Its re-evaluation of boundaries opens up questions of the social nature not only of meaning but, more fundamentally, of the mind itself. Despite its promise, however, the thesis leaves even the most uncritically accepting with a somewhat sour taste in their mouths. After all, where does extending our minds to a notebook or the Internet really get us? What new insights does accepting this paradigm enable us to discover? Putting our Popperian hats on for a minute, it seems like this kind of a hermeneutic, below the surface, is devoid of real content. Almost anything, if interpreted in the right way, can be considered part of the mind. As good scientists,

we should always be rigorously testing our hypothesis, yet, like the monumental failures of Freudian psychoanalysis or Marxist historiography, the extended mind can be applied to almost anything and, more problematically, admits of no refutation.

POSTSCRIPT—THE MEANING OF MEANING, AND AN ARGUMENT AGAINST EXTERNALISM

This discussion of the extended mind highlights an important confusion that plagues the externalist program, stemming from the definition of "meaning". Meaning is something that cannot exist, like an abstract Platonic form, outside the mind. To assert this is to commit a gross error. Every human mind is an isolated universe, in a sense. The meanings that I attach to words are my own and only my own, and they constitute what linguists call my "idiolect"[9] (the dialect that only I speak in quite the way I do). Everybody has their own idiolect that we build by trial and error from interactions with other humans using the powerful linguistic inference tools with which we already come equipped.[10] What makes communication possible is not that meaning is outside the head, but that the social nature of linguistic acquisition, together with shared common experiences (and the assumption of an intersubjective world) guarantee much overlap in the idiolects of different people who belong to what we call the same "speech community."

Notes

1. Turing, Alan M. *Computing Machinery and Intelligence.* Mind Vol. LIX. No. 236 (1950): pp. 433-460.
2. Block, Ned. "The mind as the software of the brain." In E. E. Smith & D. N. Osherson (Eds.), *Thinking: An invitation to the cognitive science* Cambridge, MA: MIT Press (1995): pp. 377–425.

3. Searle, John R. "Minds, Brain and Programs." *The Behavioural and Brain Sciences III, 3.* Cambridge University Press (1980): pp. 417-424.
4. Schank, R. C. and R.P Abelson. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures.* Erlbaum (1977).
5. Clark, Andy and David Chalmers. *The Extended Mind.* Analysis 58.1 (1998): pp. 7-19.
6. Kennedy, P.R., R.A.E. Bakay, M.M Moore, K. Adams and J. Goldwaithe. *Direct control of a computer from the human central nervous system*. Rehabilitation Engineering Vol. 8 No. 2 (2000): pp.198-202.
7. Clark, A. "Memento's revenge: The extended mind extended." *The extended mind.* Aldershot: Ashgate (2006): p. 3-4.
8. Dennett, Daniel. "True Believers: The Intentional Strategy and Why It Works." *Mind and Cognition: An Anthology.* W. Lycan (ed.) Malden, MA: Blackwell (1990): pp. 75-87.
9. E.g. see: Fries, Charles C. and Kenneth L. Pike. *Coexistent Phonemic Systems.* Language Vol. 25, No. 1 (1949): pp. 29-50.
10. E.g. see: Dominey, Peter F. *Conceptual grounding in simulation studies of language acquisition.* Evolution of communication Vol. 4, No. 1 (2000): pp. 57-85.

Works Cited

Block, Ned. "The mind as the software of the brain." In E. E. Smith & D. N. Osherson (Eds.), *Thinking: An invitation to the cognitive science* Cambridge, MA: MIT Press (1995).

Clark, Andy and David Chalmers. *The Extended Mind*. Analysis 58.1 (1998).

Clark, Andy. "Memento's revenge: The extended mind extended." *The extended mind.* Aldershot: Ashgate (2006).

Dennett, Daniel. "True Believers: The Intentional Strategy and Why It Works." *Mind and Cognition: An Anthology.* W. Lycan (ed.) Malden, MA: Blackwell (1990).

Dominey, Peter F. *Conceptual grounding in simulation studies of language acquisition.* Evolution of communication Vol. 4, No. 1 (2000).

Fries, Charles C. and Kenneth L. Pike. *Coexistent Phonemic Systems.* Language Vol. 25, No. 1 (1949).

Kennedy, P.R., R.A.E. Bakay, M.M Moore, K. Adams and J. Goldwaithe. *Direct control of a computer from the human central nervous system.* Rehabilitation Engineering Vol. 8 No. 2 (2000).

Searle, John R. "Minds, Brain and Programs." *The Behavioural and Brain Sciences III, 3.* Cambridge University Press (1980).

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

Turing, Alan M. *Computing Machinery and Intelligence.* Mind Vol. LIX. No. 236 (1950).