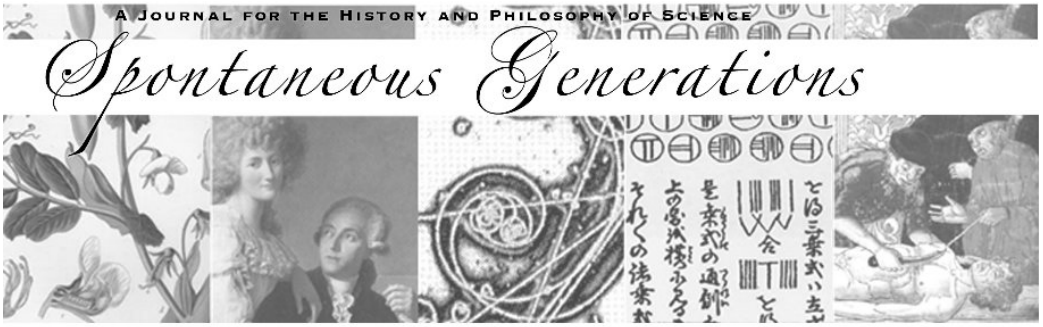

Otto in the Chinese Room

Author(s): Philip McCullough

Source: *Spontaneous Generations: A Journal for the History and Philosophy of Science*, Vol. 4, No. 1 (2010) 129-137.

Published by: The University of Toronto
DOI: 10.4245/sponge.v4i1.11718

E D I T O R I A L O F F I C E S

Institute for the History and Philosophy of Science and Technology
Room 316 Victoria College, 91 Charles Street West
Toronto, Ontario, Canada M5S 1K7
hapsat.society@utoronto.ca

Published online at jps.library.utoronto.ca/index.php/SpontaneousGenerations
ISSN 1913 0465

Otto in the Chinese Room*

Philip McCullough†

The purpose of this paper is to explore a possible resolution to one of the main objections to machine thought as propounded by Alan Turing in the imitation game that bears his name. That machines will, at some point, be able to think is the central idea of this text, a claim supported by a schema posited by Andy Clark and David Chalmers in their paper, "The Extended Mind" (1998). Their notion of active externalism is used to support, strengthen, and further what John Searle calls "the systems reply" to his objection to machine thought or strong Artificial Intelligence in his Chinese Room thought experiment. Relevant objections and replies to these objections are considered, then some conclusions about machine thought and the Turing Test are examined.

...I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.

A. M. Turing, *Computing Machinery and Intelligence* (1950)

In any case, once the hegemony of skin and skull is usurped, we may be able to see ourselves more truly as creatures of the world.

Andy Clark and David Chalmers, "The Extended Mind" (1998)

INTRODUCTION

Although his chronology may have been off slightly, Turing has proved to be a sagacious visionary. Words such as those quoted above bring machine thought closer to common acceptance. The topic of this paper began with the intent to compare and contrast content internalism and content externalism. However, upon closer scrutiny, the externalist side of

* Received 26 January 2010. Revised paper accepted 5 July 2010.

† Philip McCullough is studying at the University of Waterloo, where he began undergraduate studies in 1994 and has been studying ever since. His research interests include the importance of subjectivity to mental content in the context of the internalist/externalist debate and the philosophy of technology, and, furthermore, the role both play in cognition and personal identity.

the equation, specifically the notion of active externalism promulgated by Clark and Chalmers began to loom large in my thinking about other ideas in the philosophy of mind, including the Turing Test and its well-known refutation proposed by John Searle in the form of the Chinese Room thought experiment. Suddenly, I saw my image of the mansion of the mind in sharp relief. It was as if I had been sitting in one of the mansion's unlit rooms and a light had been switched on. I could see in a basket labelled "number one," the characters "理解." I imagined searching the Internet for this and discovering it to be the Chinese pictograph for "comprehension:" the active externalism advocated by Clark and Chalmers circumvents the argument advanced by Searle, the Chinese Room thought experiment, against the Turing Test.

Following this line of thought, I will argue that machines can (or perhaps in some sense already do) think. Whether they are cognizant depends upon how one operationalizes the definition of "mind." At the end of the previous century, two scholars proposed a definition that deviates enough from the traditional concept so as to suggest, albeit not explicitly, that certain machines already fit into the parameters delineated by this definition. I will briefly outline these ideas from Clark and Chalmers and show how their ideas regarding active externalism, by extending mind out to the environment, imply a theory that, if accepted, is devastating to one of the most important objections to the Turing Test, Searle's Chinese Room. If the mind is extended into or coupled with the environment, then the man in the room in the thought experiment does in some sense understand the story being told in Searle's parable. In this paper, I will proceed as follows: in §I, I will give a broad overview of the issues and players involved; in §II, I will detail my argumentation; in §III I examine some possible objections and give my replies to them; and finally, in §IV, I will attempt to draw some further conclusions.

Clark and Chalmers conclude in "The Extended Mind" that "there are obvious consequences for philosophical views of the mind" (1998, 18); if active externalism is a stone thrown into the reasonably calm waters of intellectual thought, this paper explores but one of the ripples it caused. There will surely be other ripples on the waters of thought, since active externalism facilitates new ways of understanding pivotal notions such as belief, information, and self-identity. Nor will these ripples reach only the abstract and deep waters of cognitive science but will continue to make themselves felt upon more distant shores, such as morality and interpersonal relationships.

I.

The central problem of the internalism/externalism dialectic is whether semantic content is at all contingent upon environmental factors. Putnam advances what has been called a passive externalist theory in his article "The Meaning of 'Meaning'" (1975). Putnam begins with an exploration of the nature of intension and extension. To this end, he presents his now famous "Twin Earth" thought experiment. He asks the reader to imagine another planet where everything is identical to Earth except that water, although perceptually identical, has a different chemical structure, XYZ, from the molecular structure of water, H_2O. Hence one cannot tell the difference conceptually to what we call water here and what is called water on the other planet ($water_2$) without reference to environmental factors since both appear exactly the same to our senses. This is meant to illustrate a position often referred to as externalism, in which certain intentional mental states are contingent upon a subject being in a certain relationship to the environment. In contrast, internalism is the notion that such mental states depend solely upon the internal constitution of an individual.

In "The Extended Mind" Clark and Chalmers postulate an alternative to the externalist/internalist dyad which they call active externalism. A thought experiment involving problem solving using a computer suggests the world is an intrinsic part of cognitive tasks, a position they wish to set apart from Putnam and others such as Tyler Burge,[1] who they label passive externalists because they only claim passive semantic aspects such as historical or distal factors make the water different from $water_2$. The authors call their third position active externalism because the relevant external semantic factors in their paradigm are actively coupled with cognition.

Furthermore, not only do cognitive processes extend into the world, the mind does as well. An example used by Clark and Chalmers is the belief that the Museum of Modern Art (MOMA) is on 53rd Street. Most people would just rely on their native memory to believe this. However, consider Otto, a person with Alzheimer's disease who is unable to remember from day to day the location of the museum; he must write it down in a notebook. Clark and Chalmers write "it seems reasonable to say that Otto believed the museum was on 53rd Street even before consulting his notebook" (Clark and Chalmers 1998, 13). He still believes that he knows where the museum is even though he must consult his notebook for this knowledge. In this manner, Otto and his notebook become a coupled system; the notebook, in this view, becomes a part of his mind. Should a twin Otto

[1] Burge, while having similar ideas to Putnam, attempts to discern a difference between "broad" and "narrow" mental content, a subtlety beyond the scope of this article.

(Twotto) on Twin Earth incorrectly jot in his notebook that MOMA is on 51st Street, then this mistaken belief will have the behavioural outcome that Twotto cannot find his destination. Hence, a propositional attitude that is part of a subject's cognition yet dependent entirely upon something outside the brain proper can potentially have a substantial impact in the physical domain. This altered conception of mind has potential effects far beyond simple geographic epistemology; one can imagine it causing changes in many other realms as well.

One such domain can be found in a foundational article in the field of cognitive science written in 1950 by A. M. Turing. In it, Turing proposes a test for artificial intelligence that is still in use today, albeit perhaps in slightly modified forms. He begins by asking "Can machines think?" In answer to this query, he proposes an "imitation game" from which roots the leaves known today as the "Turing Test" have grown. Turing was less than exact about the details of this test, on which the imitation game was based, but which is not exactly the same; the game involved identifying which of a pair of people–one man and one woman–was which. From this game, Turing went on to describe different versions of what is now commonly known as "the Turing Test"–in one version there were two entities (a human and a computer) whereas a simpler version envisions just one entity (which could alternately be human or a computer). Regardless, the quintessential idea of the game is for an interrogator to attempt to discern, by the responses to typed questions, whether the agent the interrogator is interacting with is a machine or a human. If the interrogator cannot tell the difference between the computer and a human, this is supposed to be sufficient evidence that cognition is occurring within the computer.[2]

One of the strongest objections raised against the Turing Test was proposed by John Searle and is known as the "Chinese Room" thought experiment. Searle contends that cognition occurs solely within a mind and, furthermore, the brain. The "Chinese Room" thought experiment depends upon this assumption, which I will examine in the next section. However, the thesis advanced by Clark and Chalmers undermines the conclusion that cognition can take place only in a brain that Searle draws from his now famous objection to the Turing Test. This diminishes the force the "Chinese Room" argument exerted against the Turing Test as proof of cognition.

The reluctance in some quarters to answer Turing's original question in the affirmative may stem from an anthropocentric conceit that the thought processes exhibited by *homo sapiens* must be intrinsic to and

[2] A note of gratitude to the anonymous peer reviewer whose annotations on this matter added a degree of clarity previously lacking.

inseparable from cognition, excluding by definition anything else from being the same. For centuries, humanity has privileged its cognition as special, as something that set it apart and above other forms of computational/representational processing, be they lower organic forms such as animals or, more recently, silicon-based forms such as calculators. However, this may be changing; views such as those of Clark and Chalmers are suggestive of this. Whether humanity is able to retain for itself this privileged status is an important issue, not only because it addresses foundational issues in the field of cognitive science but because it goes beyond to the very nature of mind and further yet into the realm of self-identity and what it means to be human.[3]

II.

Can machines think? Here I will examine in detail the arguments and conclusions advanced by Searle in his text *Minds, Brains, and Science* and the most salient arguments considered by Turing.

Much of the seminal paper in which Turing proposed the 'imitation game' that gave rise to today's Turing Test is devoted to an elaborate description of the parameters of his discrete state machine or computer, which seems somewhat irrelevant today, now that computers are household objects. No one would suggest he was gripping stylistically when reading these passages, but nevertheless, they are suggestive of his prescient sense of their future import. Indeed, they are indicative of the zeitgeist of the culture Turing was working in, one in which technology was seen as a panacea to humanity's ailments. Technology had just ended one of the most devastating wars in history. If it could be made to think, the possibilities must have seemed to Turing unbounded. Little wonder, then, he postulated machine thought. He then considers objections made by those answering his original question negatively. Some of these seem unworthy of serious academic speculation sixty years later, such as the theological objection or the argument from extra-sensory perception. At least one, "the head-in-the-sand" objection never was worthy, although it is indicative of a mode of thought that persists today to the point of becoming quotidian, again reinforcing Turing's uncanny ability to augur that which has come to pass. Others, however, remain problematic, such as the Lady Lovelace objection, which states that a computer is only able to produce those answers it is programmed to produce.

Like Turing, Searle is interested in the question whether machines can

[3] Self-identity in the wake of an advancing neuroscience threatening to render it meaningless is one of the most important issues in the philosophy of mind, albeit beyond the scope of this paper.

think. In *Minds, Brains, and Programs* he formulates a thought experiment, known today as the Chinese Room argument, which is still used as a critique of "strong artificial intelligence." He proposes that a computer be built that can be queried with Chinese symbols, look them up in an English rule book,[4] and return a response in other Chinese symbols in a manner sufficiently sophisticated to pass the Turing Test. This computer would convince a native Chinese speaker that it spoke Chinese. However, Searle proceeds to assert that, in theory, he could take the place of the computer, and, given access to the same rule book, could perform the same procedure. He notes that he does not actually understand a word of Chinese. Thus, a computer can mimic intelligence, but it does not have intentionality and thus can't comprehend what it is doing. Computers are merely symbol manipulators in Searle's view.

The view advanced by Clark and Chalmers addresses directly what Searle calls the "systems reply" to his Chinese Room thought experiment. After presenting his thought experiment, Searle proceeds to consider objections he believed some make in response to his intellectual offering, detailing a number of possible objections to his thesis that computers cannot think. The systems reply is of particular interest as it relates to the internalism/externalism discourse; Searle writes,

> While it is true that the individual person who is locked in the room does not understand the story, the fact is that he is merely part of a whole system, and the system does understand the story. The person has a large ledger in front of him in which are written the rules, he has a lot of scratch paper and pencils for doing calculations, he has "data banks" of sets of Chinese symbols. Now, understanding is not being ascribed to the mere individual; rather it is being ascribed to this whole system of which he is a part. (Searle 1980, 5)

If we accept the tenets of active externalism, we should also notice the similarity between Otto and his notebook and Searle and his Chinese Room and, furthermore, between Turing and his interrogator. If one accepts that Otto's notebook, upon use, becomes an extended, coupled part of his mind, then it follows that one must accept that the material upon which Searle's rules are composed, upon use, become part of his mind.

[4] It is interesting to note (as did an anonymous referee) that "tables" are often said to be used to manipulate the proper symbol/word reference when in fact Searle never uses the term; "book" and "ledger" are perhaps the closest analogue in his writing on the subject.

In the chapter of *Minds, Brains, and Programs* called "Can Computers Think?" Searle details an argument which concludes that mental states are biological phenomena. His premises are four-fold:

- Brains cause minds.
- Syntax is not sufficient for semantics.
- Computer programs are entirely defined by their formal, or syntactical, structure.
- Minds have mental contents; specifically they have semantic contents.

Searle himself concedes that the first premise is crudely formulated; he proceeds to hone his meaning to the sharper phrasing that "mental processes that we consider to constitute a mind are caused, entirely caused, by processes going on inside the brain" (Searle 1984, 39). This is true to an active externalist only if one removes that problematic caveat, "entirely caused." Active externalism allows for the causal role of factors outside the brain, such as Otto's notebook. It may also be the case that a disjunction of this premise is also true: brains or (fill in the blank here) cause minds.

Furthermore, should we adopt the active externalist paradigm, the third and fourth premises also become problematic. In this paradigm, semantic content of the mind extends beyond the traditional boundary of the head and into the environment. Otto's notebook becomes a part of his mind; likewise, the rule books in the Chinese Room become a part of Searle's mind, enabling him, when he is actively coupled with them, to understand the story. Thus if this reply is strengthened with active externalism, the Chinese Room argument loses much of its force against machine thought.

III.

A common objection to active externalism, and indeed to the idea of machine thought postulated by Turing, comes from the popular identification of cognition with consciousness. The two are commonly conflated. If it were the case that consciousness and cognition had some type of identity relationship, then it seems implausible to imbue the laptop that I type this on with consciousness, although cognition of a sort seems within the realm of possibility, if not immediately, then in the future.

To make consciousness a necessary criterion for cognition seems absurd in light of the fact that that there are many cognitive processes that are not part of our conscious thought. One example is the priming effect, well known to psychologists, where a subject asked to name a word beginning with a particular letter is far more likely to name a word used

in a sentence just prior to the task than a more common word. Whether to make consciousness a necessary criterion for mind is a more difficult question, but it is one left unaddressed by the Turing Test. Consciousness is but one small part of that bundle of processes we gather together and call "mind."

Another argument advanced against the extended mind is the criterion of portability. For something to qualify as a mind, it must be embodied and able to bring its cognitive resources to bear upon different aspects and locations of the environment. This is what keeps cognition in the head. A related idea is what might be called the "decoupling" objection to active externalism: coupling cognition with things in the environment in this manner means that decoupling them is an easy matter, and so extended cognition is not a logically necessary aspect of the cognitive process.

Portability is a more compelling objection to active externalism. Nevertheless, so long as core cognition is *reliably* coupled with external resources, this should not be an insurmountable problem. Our visual systems, for example, rely on what Clark and Chalmers call "contingent facts" about the environment to process what we see; an example is gestalt effects. So long as the coupling is stable enough, the extended mind is viable as a core cognitive process. In terms of machine thought, this is one of those tricky criteria Turing sought to avoid addressing by instead constructing a behavioural test. In this context, embodiment seems to be unnecessary.

What Searle calls the systems reply to his thought experiment entails a holistic view of the situation that arguably breaks with traditional notions of personhood. However, the idea that the entire system understands Chinese is entirely acceptable if we adopt the active externalist point of view. Searle's idea of intelligence is too narrow; active externalism allows for cognition to occur outside of brains. Again, it is the stability of the coupling that must be considered.

Finally, let us consider objections specifically related to the Chinese Room. It may be that the element of intentionality is still lacking. Many will be unwilling to grant semantic content to the Chinese Room and its inhabitant. There is a certain gravitas about Searle's aversion to the "systems reply" in "Mind, Brains and Programs"; panpsychism, the notion that mind is ubiquitous throughout the universe, potentially ensues should we accept that the Chinese Room and its inhabitant understand Chinese because we must grant other such coupled systems the same status of mind, be they Deep Blue and a chess board or a monkey using pictographs to signal hunger.

Regarding the objections involving intentionality and semantic content, it seems there are two dichotomous responses: one can accept the

Chinese Room and its inhabitant as an extended mind with semantic content or one can deny the axioms of active externalism. The notion of the nature and extent of cognition is changing; our ideas of mind must adjust with it. However, this does not entail an acceptance of traditional panpsychism; rather, new criteria for mind, such as degree of complexity and processing power, self-reflexivity and adaptability, must be established.

IV.

Machines, or computers, will at some point be able to pass the Turing Test, especially if we adopt a view of mind such as the one proposed by advocates of active externalism. Such a conceptualization accepts that cognitive processes, and indeed mind, are extended into the environment and are in many ways dependent upon it. One famous stumbling block for machine thought in the past has been Searle's "Chinese Room" thought experiment. However, no longer are computers tripped up by this argument since the extended mind obviates the Chinese Room objection to the Turing Test. The Turing Test has proven of the years to be a fertile source of intellectual innovation in the philosophy of mind. Perhaps other notions, such as those presented in "The Extended Mind," can be as fecund.

PHILIP MCCULLOUGH
University of Waterloo
pmccullo@uwaterloo.ca

REFERENCES

Burge, Tyler. 1979. Individualism and the Mental. *Midwest Studies in Philosophy* 4: 73-121.

Clark, Andy, and David Chalmers. 1998. The Extended Mind. *Analysis* 58: 10-23.

Lau, Joe, and Max Deutsch. Externalism About Mental Content. *Stanford Encyclopedia of Philosophy*. plato.stanford.edu/entries/content-externalism.

Putnam, Hilary. 1975. The Meaning of ‘'Meaning." In *Philosophical Papers, Volume 2: Mind, Language and Reality*, by Hilary Putnam, 216-71. Cambridge: Cambridge University Press.

Seager, William, and Sean Allen-Hermanson. Panpsychism. *Stanford Encyclopedia of Philosophy*. plato.stanford.edu/entries/panpsychism.

Searle, John R. 1984. Can Computers Think? In *Minds, Brains, and Science*, 28-41. Cambridge: Harvard University Press.

Searle, John R. 1980. Minds, Brains, and Programs. *Behavioural and Brain Sciences* 3: 417-57.

Turing, Alan M. 1950. Computing Machinery and Intelligence. *Mind: A Quarterly Review of Psychology and Philosophy* 59: 433-60.