The Chinese Chatroom

Separating myths, fears and intuitions from facts in a classic

John R. Searle:

Minds, brains and Programs

(and the ensuing discussion)

A project paper of

Guido Gloor

14. Oct. 2007

Abstract: *John R. Searle claimed in his text "Minds, brains and programs" and particularly the therein contained thought experiment called "the Chinese room", that there can be no such thing as strong AI.*

My intention in this paper will be to show that when Searle does so, he ultimately refers only to the intuition that we are special and that there is a secret, hidden and mystical ingredient to mental states, which he goes on to call "causal powers of the human brain". I will attempt to show how all the other claims he makes in the paper ultimately refer to this intuition, and how the intuition itself runs counter to what we know about science, machines in general and the human brain in particular.

In the end, I will attempt to show that while this doesn't prove that there can be such a thing as strong AI, at least it goes to show that Searle's attempt at proving there can't be has failed, and as such the question whether strong AI can exist is as of yet still unresolved.

1 Introduction

1.1 Disclaimer

This project paper was created during my studies at the University of Bern, Switzerland. It is not endorsed by said University, and is not an official publication in any way.

1.2 Searle's work in context

John R. Searle is a philosopher of mind, and as such, it is only natural that he's interested in artificial intelligence, whether it's possible in the first place, and if so, how and why. Well – he would be interested in the "how" and "why" if he'd believe there could be something akin to strong AI[1] in the first place, which he doesn't.

Because he doesn't just give intuitive reasons, his work does have some merit. The journal that introduced his argumentation was The Behavioral and Brain Sciences, and fortunately the nature of that journal is such that whenever an article is submitted and printed, it is first sent to selected researchers who can then give peer commentary, which in turn will be printed as well. Additionally even, the original author then is given opportunity to respond to and defend against those counterpoints and argumentations, and that answer is yet again printed alongside the original.

[1] I'll introduce the distinction between weak and strong AI according to Searle in the very next chapter. Incredibly bluntly said, strong AI is a machine that thinks, while weak AI is a machine that only seems to think.

This is a very fortunate situation, because those peer commentaries made Searle re-define some points where he was unclear. Consequently in this paper here I shall jump back and forth between his original paper and the peer commentaries where appropriate, to follow course of the logical argumentation and not necessarily the narrative continuity.

1.3 Intuitive Reasons

Coming back to the intuitive reasons I mentioned – while Searle does well in hiding them away, they come back at us like a haunting ghost, and essentially what his argumentation boils down to is: "There must be something the human mind has that a 'strong AI' doesn't".

Searle calls this something that the human mind does have and an AI can't "causal powers", yet doesn't go to explain what those causal powers could be and how they could not possibly be replicated by a Turing machine – worse, he says that even a complete replication of every single mechanism of the human brain in a computer program wouldn't be intelligent and have "real" mental states, because it would lack that secret logical ingredient. A logical ingredient that can't be replicated by a Turing complete machine, now that's something I haven't heard of before.[2]

1.4 The Structure of this Paper

I shall first (in chapter 2) give a very short account of Searle's main argumentation as to why he thinks there can be no strong AI, with a little excurse into how the Turing test works.

In chapter 3 things get interesting, when I get to Searle's secret ingredient, the "causal powers". But I'll show how there's nothing known to science that could stop a program from having those powers, even when Searle strongly seems to believe that those causal powers are special and that they can't be replicated by just[3] a machine with the right program.

Chapter 4 then will pursue the implications of that discovery, the consequences this has for the possibility of intentionality in machines. Furthermore I'll research what this means for noncognitive systems and Searle's observer-relative ascriptions of intentionality.

Finally, chapter 5 will summarize the most important points of this paper, quickly go over the line of argumentation again, and include some closing words.

---

[2] I'll expand upon this argumentation in chapter 3.2.

[3] "Just" is meant in a very quantitative sense here, although Searle would most probably include qualitative implications in this context. I will explore the intuitions that go into Searle's theory in chapter 3.1.

2 Logical Structure of Searle's Thought Experiment

2.1 Premises

Large parts of this chapter follow the structure of the text, but Searle explained quite a few things only later on. Many of the basics that he thought people knew from reading his other works when he initially wrote the essay aren't thoroughly substantiated until the author's response to the commentaries.

2.1.1 Strong AI

Searle starts his paper with a very strong assumption about what exactly strong AI is, and what it is meant to be. He will base his whole argumentation on this definition.

- The computer is not merely a tool in the study of the mind (premise)
- An appropriately programmed computer doesn't just act like a mind, but rather *is* a mind (understands, has cognitive states) (claim 1)
- Programs are psychological explanations and not just mere tools that enable us to test psychological explanations (claim 2)[4]

Three claims as it seems, however the first one is incorporated in both the following ones and thus Searle just goes to compare his findings with the ones labelled "claim 1" and "claim 2" here.

Two claims then, both of them not necessarily intuitive and widely shared. It is, however, the definition of strong AI Searle gives us (as he later in this paper proclaims[5], altering the definition won't prove his argumentation wrong, but rather disprove a different argumentation – something we'll have to agree upon), and thus we'll have to work with them.

Nevertheless I do think that some of those statements aren't shared by the majority of the defendants of strong AI. Myself I think that particularly claim 2 is unsubstantiated, as just like the human mind's physical structure alone can't be said to be a psychological

[4] See Searle 1980, p. 417f

[5] "I really have no objection to this [many mansions] reply save to say that it in effect trivializes the project of strong AI by redefining it as whatever artificially produces and explains cognition. The interest of the original claim made on behalf of artificial intelligence is that it was a precise, well defined thesis: mental processes are computational processes over formally defined elements. I have been concerned to challenge that thesis. If the claim is redefined so that it is no longer that thesis, my objections no longer apply because there is no longer a testable hypothesis for them to apply to." (Searle 1980, p. 422)

explanation of how the mind works[6], a computer program alone can't be.[7] However, even if we are to show that strong AI can't be meant the way Searle claims here, this has no further significance for disproving Searle's argument.[8]

For now, we'll have to take these claims for granted and see what Searle does with them.

2.1.2 Turing Test

The experimental setup of the Turing test works like this: A computer program and a human are both fitted into their own room. An interviewer then is given the possibility to interact with them both through a terminal.[9] If the interviewer can't tell the human from the computer after some questions, the computer is said to be intelligent.[10]

Of course, the nature, depth and complexity of these questions play e large role in how intelligent the computer seems. Obviously it has to remember things once it's told some facts, it has to be able to draw logical conclusions and has to show that it can create cross-links between new facts, common sense and its knowledge.[11]

There are many counter-arguments against the Turing test as a measure for intelligence. I don't think however that we need to look at them in-depth, as Searle may be disputing the Turing test as a reliable measure as well, but not because he thinks the test is inaccurate, but rather because he's certain that computers can't possibly be intelligent even if they do pass the Turing test – and thus that the Turing test is irrelevant in the first place.

---

[6] Otherwise there wouldn't be much left for psychology, there would just be "applied physics of the human mind". Arguably if our understanding of physics went far enough, it would be anyway.

[7] I'll elaborate in chapter 3.2.1.

[8] I'm not sure I agree – if an argument is built upon a flawed thesis, and attempts to disprove that flawed thesis, isn't the whole argument quite pointless as a whole? This is particularly true when the claims Searle makes here aren't necessarily intrinsic in his formerly noted definition of AI as a consequence of the assumption that "mental processes are computational processes over formally defined elements" (Searle 1980, p. 422).

[9] Later versions introduced other forms of communication beyond the terminal, theoretically this could be taken up to and including robotic beings that are directly confronted with the interviewer.

[10] I'll go into what this really means towards the end of this paper, in chapter 4.4.2.

[11] No simple rule-based system can reliably pass the Turing Test. Searle knows this of course, although the Chinese Room doesn't go about showing that really. I'll show why I call this "formal memory" and the implications of it in chapter 3.3.1.

Now one way to go about disproving Searle's Chinese Room Gedankenexperiment[12] would be to show that the Turing Test indeed does prove intelligence, but I'll choose another venue.[13] Searle himself admits that the Turing test could "fool" interviewers, but he doesn't give that any significance whatsoever. So he states the assumptions that followers of the Turing test might agree to:

> One of the claims […] is that when I understand a story in English, what I am doing is exactly the same – or perhaps more of the same – as what I was doing in manipulation the Chinese symbols. […] Such plausibility as the claim has derives from the supposition that we can construct a program that will have the same inputs and outputs as native speakers, and in addition we assume that speakers have some level of description where they are also instantiations of a program. On the basis of these two assumptions we assume that even if Schank's program[14] isn't the whole story about understanding, it may be part of the story. (Searle 1980, p. 418)

When he talks about this though, he states that a regular computer doesn't have the right kind of intentionality to be intelligent in the first place, and thus it's silly to even start doing tests that measure intelligence:

> I offer no a priori proof that a system of integrated circuit chips couldn't have intentionality. That is, as I say repeatedly, an empirical question. What I do argue is that in order to produce intentionality the system would have to duplicate the causal powers of the brain and that simply instantiating a formal program would not be sufficient for that.
> […]
> Now I argue at some length that they [the circuit chips] couldn't have intentionality solely in virtue of instantiating the program. Once you

[12] I'm Swiss, but me using the German word here doesn't stem from that, in fact Searle uses it as well: "Let us apply this test to the Schank program with the following *Gedankenexperiment*." (Searle 1980, p. 417, emphasis his)

[13] In the end though, we will find that the Turing test is accurate as far as accurate tests for intelligence can go. This is more a consequence of the main points of the paper though.

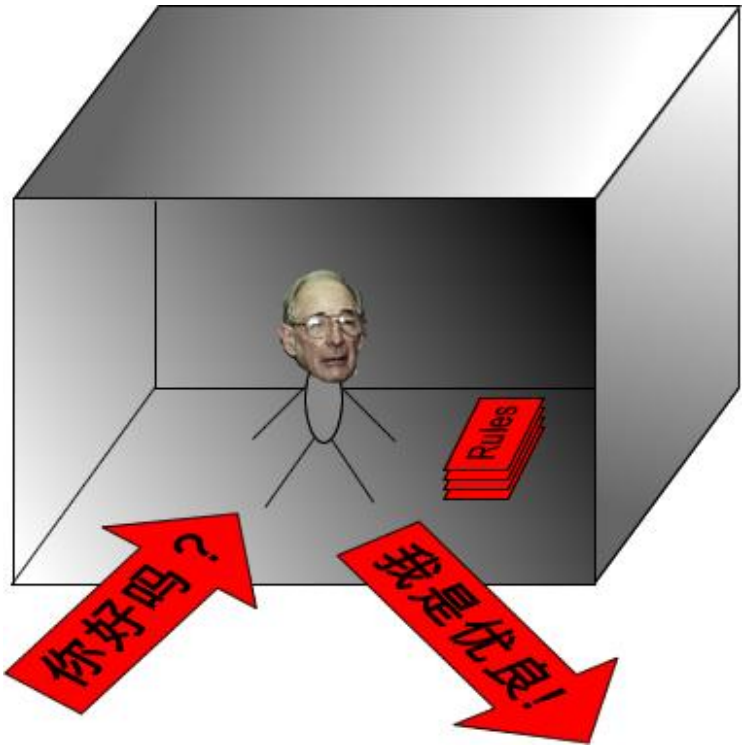
[14] Schank's program here is a program that Roger Schank created as part of his research on natural language processing (see Wikipedia 2007). Searle introduces it as follows: "I will consider the work of Roger Schank and his colleagues at Yale […] because it provides a very clear example of the sort of work I wish to examine. But nothing that follows depends upon the details of Schank's programs. The same arguments would apply to […] any Turing machine simulation of human mental phenomena." (Searle 1980, p. 417)

> see that the program doesn't necessarily add intentionality to the system, it then becomes an empirical question which kinds of systems really do have intentionality, and the condition necessary for that is that they must have causal powers equivalent to those of the brain. (Searle 1980, p. 453)

This is where I'll hook into the argument; I'll critically ask what it is that makes this "right kind" of intentionality, most of chapter 3 will revolve around this. Searle basically says that only a machine with the same causal powers as the human brain can produce any proper intentionality at all, but doesn't define what those causal powers are (they're purely process- and logic-oriented, nothing physical, although he does point out similarities between a simulation of a physical effect and a "simulation of intelligence" – we'll see more of this in chapter 4.2), and how they're not reproducible by a Turing machine.

2.2 The Chinese Room

Searle introduces his thought experiment after just one page of preface; he gets straight to the point. I'll first let a picture speak for itself, before I delve into semantics.



- There is a homunculus Searle, locked up in a room, and people keep giving him batches of Chinese symbols
- That homunculus Searle doesn't know Chinese
- He has a set of rules in English, which he understands
- He can identify symbols by their shapes (purely formal symbols)

- With that set of rules of his, he can do several things:
    - correlate symbols with one another
    - correlate symbols of one batch with those of another batch
    - produce new (formal) symbols and hand them back to the people who gave him the original batches

The batches he's given by the people outside the locked room have names and meanings, for them, albeit not for him:

- A first batch, the "script"
- A second batch, the "story"
- A third batch, the "questions"
- Furthermore, the rules are called "program"

For him, the squiggles don't have any meaning at all. And even if he can infer from being handed three batches of English text later on that the three squiggly batches here have the same significance for the experiment, this isn't important for the experiment: Searle says that a human learning from following the rules he's following isn't understanding thanks to the rules, but rather learning thanks to them, which is not the same thing at all, and again we'll have to agree:

> Suppose that the program contains such instructions as the following: when somebody holds up the squiggle-squiggle sign, pass him the salt. With such instructions it wouldn't take one long to figure out that "squiggle squiggle" probably means pass the salt. But now the agent is starting to learn Chinese from following the program. (Searle 1980, p. 451)

For those people outside giving him these things they call „questions", his output is indistinguishable from the one of a real Chinese speaker. This of course means that the rules are hugely complex, but there's another thing I shall get back to – it also means that the rules have to have indefinitely many cases covered, or they have to have some kind of state preservation mechanism, like, have Searle remember "squiggle squaggle" and later act

according to not only the new input (the questions), but also according to that internal preservation mechanism. Let's just call it "formal memory" for now.[15]

2.2.1 Searle's Posits

Now the thought experiment goes on and Searle has the people outside hand an English text to the homunculus Searle, and have him hand back out English responses to the questions raised in there. Now Searle posits the following things:

- That the responses in English are as good as the ones in Chinese and vice versa
- That the homunculus Searle produced both kinds of answers[16]
- That the homunculus Searle did understand the English text and answered the English questions
- That the homunculus Searle did not understand the Chinese text, didn't even realize he seemed to answer questions and just did formal symbol manipulation with squiggles when he seemed to answer there[17]
- That what the homunculus Searle represents is akin to being the instantiation of a computer program[18]

2.2.2 Searle's Conclusions

Searle then gets back to the original position and tries to show how the two claims of strong AI aren't satisfied here: Neither does the homunculus Searle understand, although he

---

[15] This will be important in my discussion of how the system of rules has to be built, in chapter 3.3. I will not, however, go into how exactly that formal memory has to be built or how it's instantiated in our brains – that's not really a matter for philosophers to solve but rather one for neurobiologists, and maybe neurophysiologists. I will however give a proper definition of formal memory in chapter 3.3.1.

[16] This does seem overly blunt to me. Searle really is convinced of this though when he reiterates: "[…] my answers to the questions are absolutely indistinguishable from those of native Chinese speakers." (Searle 1980, p. 418)

[17] I don't intend to disagree with this at all.

[18] I don't intend to agree with this at all. Rather, what my intention will be is to show that Searle here is something akin to the CPU in a computer, and even if there is such a thing as "strong AI" nobody will seriously say that it's the CPU that does the understanding. If we'd have to fit this into Searle's response categories, it would probably be some kind of systems reply. Basically, on one hand, the homunculus Searle is the origin of the English answers, while on the other hand the homunculus Searle is the means of the "system of rules" (including the formal memory, I guess) that produces the Chinese answers.

certainly seems understanding enough to the outside world (which would be claim 1).[19] Nor does the program, the system or the homunculus Searle serve to explain anything about understanding or the human mind (claim 2). However, here he has to refer to intuition for the first time:

> I have not demonstrated that this claim is false [that when I understand a story in English, what I am doing is exactly the same – or perhaps more of the same – as what I was doing in manipulating the Chinese symbols], but it would certainly appear an incredible claim in the example.[20] (Searle 1980, p. 418)

Searle then comes to a very central point in his argumentation:

> […] what is suggested […] by the example is that the computer program is simply irrelevant to my understanding of the story. In the Chinese case I have everything that artificial intelligence can put into me by way of a program, and I understand nothing; in the English case I understand everything, and there is so far no reason at all to suppose that my understanding has anything to do with computer programs […]. (Searle 1980, p. 418)

So what he is saying here is the following: The homunculus Searle doesn't understand a Chinese text, even though he is the one who follows the rules (or the "program"). And even though he doesn't follow the "program", he understands an English text. Digging deeper, what Searle implies is that it has to be homunculus Searle that understands in both instances. So essentially, Searle's message is:

Only if the homunculus Searle understands, there is understanding. Or, in other words: If the homunculus Searle doesn't understand, nothing and nobody does – even when both the rules he follows and the formal memory that keeps track of states are external to him.

[19] Our reservations concerning whether it really is the homunculus Searle who has to do the understanding still apply of course, we're following Searle's argumentation here still though.

[20] I will expand upon why I think the quoted position doesn't hold in chapter 3.

3 Searle's Mysterious Causal Powers

3.1 Searle's openly admitted Intuitions

Searle commits in quite some places to the view that believing in strong AI can't be real, is silly, strange and borderline idiotic. I'd like to present you a collection of quotes that illustrate this, and that might help illuminate the motivation behind the whole thought experiment.[21]

So if I am to go into what it is that makes those arguments have some weight, I'll have to look into where they're coming from first. Searle makes this quite clear, although he never explicitly tells us the foundation of his argumentation. I want to lay those fundaments bare, and I'll start by showing what the intuitions Searle bases his argumentation upon are, commenting every one of the passages I found to that end.

> Now I find the thesis of strong AI incredible in every sense of the word. (Searle 1980, p. 450)

This is the introductory sentence to a paragraph that goes on like this: "But it is not enough to find a thesis incredible, one has to have an argument, and I offer an argument that is very simple: […]". Still, it shows how the idea that strong AI can't be real came first and the argument came later. Of course, being motivated by intuition is nothing bad at all, but letting intuition blind one's view is bad, and the purpose of this paper is to find out if Searle did let just that happen.

> Well I don't know *how* the brain produces mental phenomena, and apparently no one else does either, but *that* it produces mental phenomena and that the internal operations of the brain are causally sufficient for the phenomena is fairly evident from what we do know. (Searle 1980, p. 452, emphasis his)

This in and of itself is an intuition. It correlates closely with the main point of Searle's position: There is something special about states of mind, and "we know there is", thus there is. However, this requires a belief in the intuition that we are special, a thing I don't

[21] Keep in mind that this chapter is there mainly for illustrative purposes and that I mean no disrespect whatsoever towards one of the greatest philosophers of mind of our time. However, as it happens, I disagree with both some of the premises and the conclusions he presents in his paper, and my intention is to present the differences that lie at the very core of that dissent.

have in this context.[22] At another position, he even goes into details as to what kind of special we are:

> What I do argue is that in order to produce intentionality the system would have to duplicate the causal powers of the brain and that simply instantiating a formal program would not be sufficient for that. (Searle 1980, p. 453)

So, he thinks we're special in that our brain has some causal powers that a computer doesn't have. As I will illustrate later (in chapter 3.2.1), Searle at the same time believes that those causal powers are purely logical and that a Turing machine can't reproduce them. But I'll let that rest until then, back to Searle's motivations:

> On the position of strong AI there cannot be any empirical questions about the electrochemical bases necessary for intentionality since any substance whatever is sufficient for intentionality if it has the right program. I am simply trying to lay bare for all to see the full preposterousness of that view. (Searle 1980, p. 453)

Well, I don't see anything preposterous in that. Maybe I'm not religious enough. The strongest of Searle's assumptions shows again here:

> But I am not in any sense looking for a criterion for the mental. I know what mental states are, at least in part, by myself being a system of mental states. (Searle 1980, p. 455)

Searle is smarter than me. I have no clue what mental states are, as chapter 3.3.2 shows. I don't believe that mental states have anything special about them, or that they have any true intentionality[23] attached to them that other, non-mental, states couldn't possibly have.

His response to people calling him out about that whole system of assumptions that he bases the argument upon is the following, completely dodging the issue (because the intuitions people called him out upon were never about whether the homunculus Searle understands in and of itself):

> But consider. When I now say that I at this moment do not understand Chinese, that claim does not merely record an intuition of mine, something I find myself inclined to say. It is a plain fact about

[22] Actually, I don't have any belief whatsoever that we're special in any way, shape or form, but that's not the issue – the point is that Searle's argumentation seems to require the reader to share this belief.

[23] I will dissect this in chapter 4.4, what I'm talking about here is Searle's *intrinsic intentionality*.

> me that I don't understand Chinese. Furthermore, in a situation in which I am given a set of rules for manipulating uninterpreted Chinese symbols […] it is still a fact about me that I do not understand Chinese. (Searle 1980, p. 451)

In fact, I quite agree with Searle on this issue (and it's hard not to really); the homunculus Searle does not understand Chinese. But arguably, that has nothing to do with the intuitions many critics were speaking about. Searle takes the easy way out here and doesn't challenge his beliefs about whether there is something special about our mental states to start with.

I strongly believe that the question whether the homunculus Searle understands anything has nothing to do with whether there is anything that can be called "understanding" in the whole Chinese Room (including rules, states, decisions, executions and input/output) – after all, everything he does is execute the formal rules, he's the CPU in a computer, the read-write head in a Turing machine, and doesn't even need any intentionality for that in the first place. He just needs to know how to follow directions.

3.2 A Turing Machine's Powers

If we want to find out whether Searle's argumentation holds, we'll have to introduce Turing machines. Because when he says "a computer with the right program can't have the right kind of causal powers", he's saying that "a universal Turing machine can't have the right kind of causal powers".

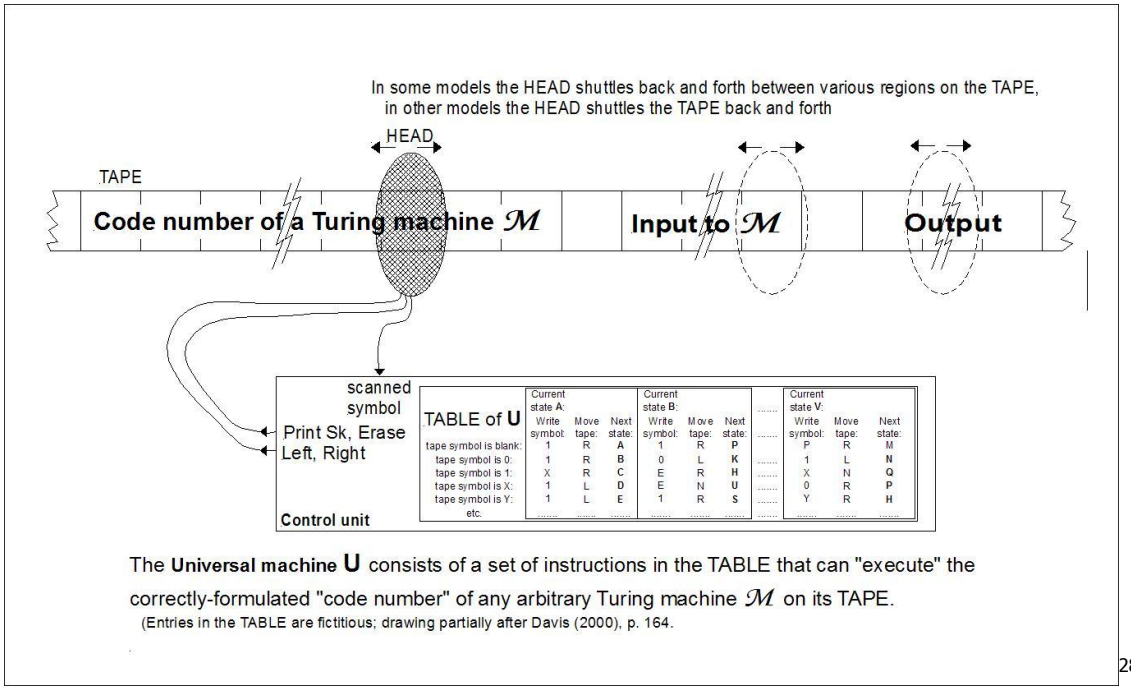
This is true because it's an abstraction: Because the universal Turing machine is the theoretical model for *any* computer with *any* kind of program and thus a specific combination of computer and program is always a Turing machine, we basically clear up the argument and remove unnecessary fluff – we no longer need to think about whether the program we're using is the right one, and we can forget about computers and their physical limits as well. Since Searle didn't go and specify that the computer's processing units have to be built from circuit chips,[24] and nothing about how exactly the AI has to be built,[25] we're not losing any information either.

---

[24] Actually, he does implicitly specify it, as whenever he talks about the workings of a computer he does talk about circuit chips. But it's not really important to him: "The circuit chips in his [Pylyshyn's] example would necessarily have intentionality, and it wouldn't matter if they were circuit chips or water pipes or paper clips, provided they instantiated the program." (Searle 1980, p. 453)

[25] And rightly so, he does want to prove that there can be no such thing as strong AI in the first place, independent of how the implementations improve. Thus by the way the reason why you won't read anything about current AI research in this paper – that wouldn't add anything of importance to the argument here.

So, what is a Turing machine? Basically, it's an abstraction of a stateful computational device. We have a tape, and a head. The head can read and write on the (one-dimensional) tape, and it can move back and forth.[26] The machine then also has a system of rules according to which (in combination with the last scanned symbol) it moves the head[27] and reads and modifies data under the head.



[28]

So far so good, but the question arises: Why is this machine interesting in the first place? The answer to this is simple, and I hinted at it already some paragraphs ago: Because it's a model for *every* machine. [29] Everything that can be expressed in any machine can be

[26] Or, it can move the tape back and forth. This is mostly a matter of reference system of course.

[27] Note the similarity, the read-write head here does pretty much what the homunculus Searle does in the Chinese room. One could even say that the head gets the rules handed and acts according to them, but that would be ascribing intentionality to it, and we don't want to do that just yet, do we?

[28] Picture courtesy of Wikipedia author Wvbailey, covered by the GNU Free Documentation License 1.2 (Wikipedia 2007)

[29] At least, it is a model for every single-tasking machine, having serial computations. However, there are several things that profoundly alleviate the problem this seems to bring up: The brain is massively parallel. Those things are: We do have multitasking, which basically breaks up parallel tasks into chunks and processes them quasi-parallel in a sequential manner (switching from task to task all the time). There are meanwhile multi-core processors, parallel calculations are frequently done in computer networks too, so we also have (somewhat limited, physically) real parallel calculations even, which brings us multiple parallel Turing machines. Also, if we wanted to simulate the behaviour of neurons themselves we could always use calculus and advanced physics and mathematics. So in the end, the Turing machine model works for parallel calculations as well.

expressed in a universal Turing machine – this is known as the Church-Turing thesis[30] and admittedly, it is not mathematically proven. However, no counter-example has been found so far, so the theory hasn't been falsified either, and the winds of doctrine[31] tell us that nobody seriously doubts the truth of it. [32]

A computer basically is a finite implementation of a universal Turing machine. This leads us to the one constraint a computer has, compared to the machine: A computer can't calculate everything a universal Turing machine can. Simply because a Turing machine has infinite resources, while a computer doesn't. However, a brain doesn't have infinite resources either, and we are presuming an adequately powerful computer for strong AI, so this constraint doesn't have any significance for the Chinese room discussion.

3.2.1 About Causal Powers a Turing Machine Can't Have

This bit is interesting. Because, as we've seen in the last chapter; there are no computable problems that a (universal) Turing machine can't solve. I will argue that causal powers are logical entities, thus per definition computable, and thus there can't be any kind of causal powers a Turing machine can't replicate.[33] Note that causal powers are used in a very strict sense here, but Searle explains this best himself:

> […] I do not, of course, think that intentionality is a fluid. […] I think, on the contrary, that intentional states, processes, and events

[30] The Church-Turing thesis is also known as Church's thesis, Church's conjecture and Turing's thesis; see Wikipedia 2007. The thesis is, essentially: "What is effectively calculable is computable" (Gandy 1980) – or, in many more but less unclear words, "What is produced by any means whatever that uses a method each step of which is precisely determined and which is certain to produce the answer in a finite number of steps, can be produced by a Turing-machine or equivalent mechanical device." (Compiled from the following: Wikipedia 2007 and Rosser 1939).

[31] These are the things Searle refers to when he says: "[…] intentional states, processes, and events […] are [necessarily] both caused by and realized in the structure of the brain. Dennett assures me that such a view runs counter to 'the prevailing winds of doctrine.' So much the worse for the prevailing winds." (Searle 1980, p. 451) – I'll provide a partial answer as to why that argumentation doesn't hold in chapter 3.2.2, but it does seem a bit ignorant to just, well, ignore large parts of science because they provide counter-arguments to a theory when attempting to prove that very theory.

[32] Of course, just like every other theory, the Church-Turing thesis is prone to eventual falsification, or it would have to be adapted to include special cases where it would not apply. However, so far no reason to do so has arisen, and the theory is quite well-established. Challenging one of the fundamental pillars of computability theory better had a better reason than "we do have got to be special, don't we?"

[33] It is important that we're talking about replication and not simulation here, I'll explain in chapter 4.2.

> are precisely that; states, processes, and events. The point is that they are both caused by and realized in the structure of the brain. (Searle 1980, p. 451)[34]

We know that there are things a Turing machine can't compute. That are, namely, undecidable (or, a word I prefer, incomputable) problems like the halting problem.[35] So essentially, the question whether there can be causal powers that a Turing machine can't have boils down to the one that will occupy us in chapter 3.2.2: Are the causal powers of the brain incomputable?

So far though, a different question arises: Does strong AI really want to provide an explanation for the human brain? I said in the introduction already that I don't agree with Searle's proposition two, and here's why: A Turing machine can't do everything. Consequently, strong AI can't do everything. Namely, a Turing machine can't necessarily explain another Turing machine – that's the very essence of the aforementioned halting problem. So, just like analyzing a human brain with a human brain won't yield a complete explanation, neither analyzing a strong AI[36] with a human brain will yield a complete explanation, nor will analyzing a human brain with a strong AI (or a strong AI with another strong AI) do so – we'd need a super-strong AI, a thing much more powerful than the human brain, to get there. And no reputable computer scientist will want to prove the opposite.

So while we do work with the claims of strong AI that Searle proposed, they do seem unrealistic. Still I think this paper will go to prove that even with those unrealistic claims Searle doesn't have a serious counter-argument to the potential existence of strong AI.

3.2.2 Are the Causal Powers of a Human Brain Incomputable?

We found out in the last chapter that this question really is important, because essentially Searle claims that a Turing machine can't have those causal powers and thus (because everything that is computable can be replicated by a Turing machine) they have to be incomputable. However, he softens up that argument a bit later on:

> Some of the commentators seem to suppose that I take the causal powers of the brain by themselves to be an argument against strong AI. But that is a misunderstanding. It is an empirical question whether any given machine has causal powers equivalent to the

[34] He also says that "at every level, the phenomena are causal." (Searle 1980, p. 455)

[35] See Wikipedia 2007.

[36] For now, assuming there is such a thing.

> brain. My argument against strong AI is that instantiating a program is not enough to guarantee that it has those causal powers. (Searle 1980, p. 452)

Still, Searle seems to be tilting at windmills here - if Searle was to say that the causal powers he's talking about can't be computed, he would have to show why that is. The burden of proof lies upon him, not us. On the other hand, if he says that it's just the current implementations of AI that don't have the causal powers but in theory, it would be possible to build a strong AI with a Turing machine, he's contradicting himself. So Searle's attempt at softening up his own argumentation either destroys it or doesn't really soften it up in the first place – still it's impossible for a computer (even with the right kind of program) to have those causal powers, or Searle doesn't have an argument at all.

I might have to add a little digression concerning scientific theories: I'm a strong follower of W. Quine in this matter. I think the core of the Chinese Room argument is pretty similar to the one Quine introduces when he talks about how scientific theories relate to each other:

> As an empiricist I continue to think of the conceptual scheme of science as a tool, ultimately, for predicting future experience in the light of past experience. Physical objects are conceptually imported into the situation as convenient intermediaries – not by definition in terms of experience, but simply as irreducible posits comparable, epistemologically, to the gods of Homer. Let me interject that for my part I do, qua lay physicist, believe in physical objects and not in Homer's gods; and I consider it a scientific error to believe otherwise. But in point of epistemological footing the physical objects and the gods differ only in degree and not in kind. (Quine 1951)

Similarly, I don't believe in Searle's causal powers but rather said strong winds of doctrine, simply because they allow me to predict more of the future more adequately, and they're simpler (as theories) than additional mythical and inexplicable[37] causal powers. Consequently, if the causal powers were incomputable and I would be offered some kind of

[37] And they are inexplicable, Searle tells us: "[…] it then becomes an empirical question which kinds of systems really do have intentionality, and the condition necessary for that is that they must have causal powers equivalent to those of the brain." (Searle 1980, p. 453) So, by conclusion, the only way to find out if something has these causal powers (and they are the thing defining intentionality in the first place, as both necessary and sufficient criterion, so looking for them is equivalent to looking for intentionality) is by … finding out it has, which is impossible by the simple means of a Turing test or similar behaviourist methods, or any other means bar intuition. Or so it seems.

proof to that end, I'd gladly accept. Just taking them (including their alleged incomputability) for granted however won't do.

Additionally, I haven't found out yet how Searle would classify something as having or not having the causal powers, but he defines: "I think it is evident that all sorts of substances in the world, like water pipes and toilet paper, are going to lack these powers, but that is an empirical claim on my part." (Searle 1980, p. 453) Yes, it is an empirical claim; however it doesn't seem to be a testable one and as such, it's pure doctrine.[38]

3.3 The System of Rules

The system of rules Searle proclaims has to be immensely complex. Let's just take a simple example: The researcher tells its subject during a Turing test the following:

> For the rest of the discussion, we will call all lamps sheep, and all light switches we shall call trees. Furthermore, we shall substitute 'use' with 'eat'. So, if I switch on the light, did my eating the tree change anything about the sheep in my room?

The rules have to keep track that several words are now used differently from their usual meanings, they have to remember facts dispersed throughout the conversation, they have to keep track of who's asking and who's answering – many of these things are meanwhile possible, but a system of rules that satisfies all these points and is able to answer "yes, the sheep are lit up now" hasn't been created yet.[39]

This complexity is something some critics think suffices for disproving Searle's theory – he seems to largely underestimate the complexity of such a rule system. For example Robert P. Abelson raises this objection in his peer commentary:

> First of all, it is no trivial matter to write rules to transform the "Chinese symbols" of a story text into the "Chinese symbols" of appropriate answers to questions about the story. To dismiss this programming feat as mere rule mongering is like downgrading a good piece of literature as something that British Museum monkeys can eventually produce. (Abelson in Searle 1980, p. 424)

[38] Myself I don't see how neurons would be any more probable as producers of intentionality as well, they don't seem much more reasonable than water pipes – if we wouldn't "know" they produce Searle's *intrinsic intentionality*, we'd never guess.

[39] Even if there was a program that could give the correct answer here, we'd certainly find another, yet more complex example that can't be answered by a machine correctly as of yet. When and if we do find or program true "strong AI", that would of course change.

While that it seem that what Searle does is exactly dismissing that programming feat as a trivial matter of easily created rules, it's not certain he does, and it doesn't seem to be the core issue with his argument to me. Searle himself, in his answer to this comment, goes one step further and proclaims:

> […] it is no mean feat to program computers that can simulate story understanding. But, to repeat, that is an achievement of what I call weak AI, and I would enthusiastically applaud it. […] I am afraid that neither this nor his other points meets my arguments to show that […] there is no reason to suppose that instantiating a formal program in the way a computer does is any reason *at all* for ascribing intentionality to it. (Searle 1980, p. 453f, emphasis his)

So it all goes back to the question whether we have any reason for ascribing intentionality to the computer-program-system in the first place. But I'd like to follow this thought a bit further and show what the basic requirements of such rules would be.

3.3.1 Formal Memory – Definitions

Already when I talked about the Chinese room thought experiment in chapter 2.2, I introduced this rather abstract term for all the state keeping a system of rules has to have. Now, what is formal memory? I'd like to first find that out[40] and then, in the next chapter, find differences between formal memory states and mental states.

First off though, we have to ask if formal memory really is necessary. Can't we just have a very wide tree of rules, where we just have one (the current) position and a wide array of choices from there onwards, and we don't have to backtrack to former information for knowing how to proceed at any given point?

Well, yes, we can,[41] but that won't rid us from the necessity of states, albeit only one kind of state: Our current position in that decision tree. Since formal memory is state keeping and we do need states, we also need state keeping and thus we have formal memory. Because having one large tree doesn't rid us from the necessity of formal memory anyway, I'll just

[40] Actually, it will be as much a definition as it will be finding out what it reasonably is.

[41] Such an implementation would however be incredibly inefficient. If the rules necessary to duplicate proper understanding with good state keeping are immense, the ones without said state keeping are exponentially more complicated. And since they'd involve lots and lots of duplicated bits and pieces too, any programmer would want to introduce shared substructures. But we're getting off topic.

assume that every state that should reasonably be isolated[42] will be kept separately in that formal memory we're about to introduce.

Formal memory then finally is literally any way to keep states, and link them together.[43] Formal memory in the sense we're using it here does have to keep track of facts about the outside world, it has to keep track of goals and mark them as such, it has to be dynamic and flexible so it can adapt to ad hoc definitions and new inputs, and it has to be able to correlate one fact to numerous others in various different semantic contexts, presenting related (formal) states on the fly. Formal then means that no true, real, semantic meaning is attached to them.[44]

But what does that mean, at its very core? Just one thing: Formal memory has to be able to keep track of data, it needs rules governing how it keeps track of that data, and it needs some means to store that data. Later I'd like to explore the differences between generic ways of keeping data (or formal states) and mental states. But first we have to define "states":

States here are used in a rather abstract sense, anything that represents a single matter of fact or a memory is a state, no matter how it's implemented. For example, a goal or memory in the human mind. But just as well a couple of nodes simulating neurons holding specific parameters might be the thing that holds the state, or it might be a series of zeroes and ones that (8 bits at a time) represent characters which are then to be further processed, or a series of logical "if-then-else" constructs that leads to the correct composition of substates. On the physical level then the state might be stored in (real) neurons and their biologic properties, in a couple of bits and bytes on a hard drive (magnetic orientations in sub-surfaces of physical media of varying characteristics), or bits and bytes on a CD (pits and lands in a layer of aluminium on a plastic surface), or on a RAM chip, or through any other means of storing data.[45]

---

[42] And I'm talking about all kinds of states here. Facts, "beliefs", "goals" – and I intend to get rid of those quotes around the latter two too, but it's a bit early for that still.

[43] I really want to introduce formal memory as a very broad and unconfined matter, within reasonable borders of what a (size-limited) Turing machine can do. This is because every possible kind of formal memory is thinkable, and for strong AI we have to assume that we do have the right kind of program and as such, the right kind of formal memory handling.

[44] As I intend to show in chapter 4.4 however, I doubt that there is such a thing as Searle's intrinsic intentionality or the true semantic meaning we're talking about here.

[45] Although I'm only speaking of storing non-mental data in bits and bytes here, that's of course not the only way to do this. Since we're independent of the physical implementation of the storage algorithm though, that can safely be ignored.

The point is, those implementation details on various levels are completely irrelevant for our definition of formal memory and (mental or non-mental) states, yet it's important to note that whatever the details, the state is considered to be the same if it represents exactly the same fact.[46] So states are a purely logical construct, aiding us to abstract from the physical and to concentrate on what things are really about. Or, said differently, states are an abstraction that lets us forget about (in the context) unnecessary details.

3.3.2 Just States: The Difference between Mental States and Non-Mental Ones

Searle wants us to follow his argumentation by introducing a profound difference between cases where "understanding" applies, and cases where it doesn't – or, in other words, cases where states are mental and cases where states are not mental:[47]

> My critics point out that there are many different degrees of understanding; that "understanding" is not a simple two-place predicate; […] that in many cases it is a matter for decision and not a simple matter of fact whether x understands y, and so on. To all these points I want to say: of course, of course. But they have nothing to do with the points at issue. There are clear cases in which "understanding" literally applies and clear cases in which it does not apply; and these two sorts of cases are all I need for this argument. (Searle 1980, p. 418f)

Arguably, they are the only things he does need for his argument, but my question will be if there's anything that makes mental states that special in the first place. If they're not special at all, the question whether "understanding" applies must be an entirely different one than the one about the existence of mental states.

---

[46] This does however imply that in every non-trivial case, implementation details will make representing exactly the same state in different implementations quite impossible. However, when we'll talk about mental states later we have to keep in mind that no two brain structures are exactly the same, and as such it won't happen that two brains, or even the same brain at different times, will hold exactly the same contents – and seeing how states are built in a way that they interact with each other too, this makes having exactly the same mental state twice (even within the same brain) highly improbable. So the difference between brains, so to say, is not that much smaller than the difference between brains and other means of storing states. This is a vivid and very interesting field of research of course; see Sehon 2005 for a discussion of the problems this brings to both common-sense psychology and "strong naturalism", as he calls it.

[47] Of course, the latter doesn't necessarily mean the same as the former, which will be my whole point in the end. But if I understood him correctly, according to Searle's nomenclature they are the same.

I'd like to start my counter-argument with Searle's definition he (reluctantly) gives of what mental states are:

> Mental states are as real as any other biological phenomena. They are both caused by and realized in the brain. That is no more mysterious than the fact that such properties as the elasticity and puncture resistance of an inflated car tire are both caused by and realized in its microstructure. Of course, this does not imply that mental states are ascribable to individual neurons, any more than the properties at the level of the tire are ascribable to individual electrons. To pursue the analogy: the brain operates causally both at the level of particles and at the level of its overall properties. Mental states are no more epiphenomenal than are the elasticity and puncture resistance of an inflated tire, and interactions can be described both at the higher and lower levels, just as in the analogous case of the tire. (Searle 1980, p. 455)

Funny enough, I agree in large parts with this analogy. Mental states are something the brain has much like elasticity[48] is something a tire has. The exact nature of *our* mental states is profoundly influenced by the structure of our brain. However, the reason why I find myself agreeing so much is one Searle can't have intended: Elasticity isn't restricted to tires only. Rubber has elasticity, horn has elasticity,[49] wood has elasticity (albeit drastically lower); even metal has elasticity (although that one is a lot lower yet again). We can even say: every object constructed from any material has elasticity, as long as it's in its solid state.[50] What the material that the object is made of influences is not the fact that there is such a thing as elasticity; rather the material influences the amount, extent and maybe direction of that elasticity.

Furthermore, not only materials can be elastic, but structures as well. A spring for example has elasticity, and a completely different elasticity from the metal it's made of (although the two are, of course, linked).

This is because elasticity is not a physical property of molecules or even clumps or structures of molecules per se, and the presence of elasticity is not dependent on there

---

[48] For simplicity, I'll refer only to elasticity from here onwards. A very similar argument could be made for puncture resistance, or colour, or much other behaviour that materials only exhibit once we pass beyond the molecular level.

[49] Just think how elastic our hair is.

[50] Even this limitation isn't a necessary one; some materials exhibit elasticity even when they're in their fluid state.

being specific properties of the materials we're talking about. Rather, elasticity is an abstraction; it is a virtual property, a way of talking about similar behaviours in completely different materials. It is not a matter of empirical tests whether any given material has elasticity; it's just the amount of elasticity a material has that is in question and subject to said empirical tests.

What does this have to do with the difference between mental and non-mental states? And how does this fuel my forthcoming argument that there is just one kind of states, and they're not decidedly mental or non-mental?

If elasticity is something all materials have, mentality (being very much alike in nature, considering both are abstractions and ways of talking about behaviour materials and compounds of materials only exhibit once we leave the microscopic space) for states must be similar. Unless we find something that mentality specifically has and elasticity doesn't. Searle's proposition as to what that something is that makes mentality special are causal powers that the brain is able to produce but normal Turing machines not; as I've shown in chapter 3.2.2 the postulation of such powers is pure doctrine and runs counter to what we think we know about computability. I'm not a big fan of unsubstantiated and esoteric claims;[51] as such I'd like to deny those mystical causal powers any argumentative power unless they're proven and shown with more than pointers to the intuition that "we have got to be special".

Considering we now exposed the difference between mental states and non-mental states as a myth, I'd like to introduce a somewhat novel concept, just[52] states: They're just that, states. Not mental or non-mental ones, just states. Mentality is an attribute that we can give them, we can even give them different amounts of mentality based upon our intuition, but there's nothing fundamental or physical[53] that separates mental from non-mental states.

3.4 The Systems Reply, and Searle's Answer

Searle provides quite a few answers to commonly raised objections to his theory already in his main paper, and more answers later upon the response to the peer commentaries. Since the answer I present in this paper is most closely related to the systems replies Searle mentions, I'd like to show Searle's answers to these too, and where they fail when it comes to my argument.

---

[51] Particularly when said claims both are counter-intuitive (for me) in the first place, and when the conclusions they have for somewhat unrelated subjects are counter-intuitive as well, admittedly.

[52] Just as in "just so", not "justice".

[53] Apart from the fact that states "caused by and realized in the brain", as Searle calls them, are the ones that most often have mentality attributed when people speak about them.

Searle gives this definition of the systems reply, paraphrasing the people promoting it:

> "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, p. 419, quotes in original)

He then goes on to give a reply, proposing that the homunculus just should internalize all the aspects of the system, including rules and data banks and calculations, but he sees that a follower of the systems theory will just go on to say that this essentially just leads to two subsystems within the homunculus become man, one of them understanding English and one of them understanding Chinese, and they have little to do with each other.[54] Searle then replies:

> But, I want to reply, not only do they have little to do with each other, they are not even remotely alike. The subsystem that understands English […] knows that the stories are about restaurants and eating hamburgers, he knows that he is being asked questions about restaurants and that he is answering questions as best as he can […], and so on. But the Chinese system knows none of this. Whereas the English subsystem knows that "hamburgers" refers to hamburgers, the Chinese subsystem knows only that "squiggle squiggle" is followed by "squoggle squoggle." (Searle 1980, p. 419)

It seems that we will have to look at what "knowledge" is. In the sense that Searle is using it here, it's relating one fact to the other. Three kinds of knowledge are required here: what a hamburger is, what a restaurant is, and what questions and answers are. However, all those things can well be facts in our now introduced formal memory. All the things about texture, taste, look, and generally context that a human being knows[55] can just as well be

[54] Searle introduces this "extended systems reply" only after he refutes the original reply. This is due to narrative structure mainly I guess.

[55] There's other things too that aren't even intersubjective (or, even less intersubjective than the attributes already mentioned). For instance, one person might like hamburgers while another one doesn't, for one person the first intuition with hamburgers is Krusty Burgers while another one thinks of Burger Queen first, one attributes "tastes good" and another one "makes fat" as a first thought –

represented in a formal memory[56] and so what differentiates "knowledge" from just states is, again, attribution of the mental.

So as a matter of fact, chunks of knowledge are just states. When they're instantiated in formal memory, they're usually called "non-mental", while when they're instantiated in the human brain, they're usually called "mental".

Searle introduces something else, "semantic content", when he later replies to commentaries that put Searle's argumentation about the systems reply in question:

> Both Wilensky and McCarthy fail [in their commentaries] to answer the three objections I made to the systems reply.
> 1. The Chinese subsystem still attaches no semantic content whatever to the formal tokens. The English subsystem knows that "hamburger" means hamburger. The Chinese subsystem knows only that squiggle squiggle is followed by squoggle squoggle.
> 2. The systems reply is totally unmotivated. Its only motivation is the Turing test, and to appeal to that is precisely to beg the question by assuming what is in dispute.
> 3. The systems reply has the consequence that all sorts of systematic input-output relations (for example, digestion) would have to count as understanding, since they warrant as much observer-relative ascription of intentionality as does the Chinese subsystem. (Searle 1980, p. 453)

When we realized that knowledge is mental states, and thus just states that we are inclined to call mental, the semantic context Searle mentions in the first of these points gets put into perspective as well. We introduced formal memory[57] as a thing that is dynamic, adaptable and has links from one facet of a fact (or state)[58] to another. So when the brain has various crosslinks to other facts that make us think of the fact having semantic content, formal memory does too. Unless we find Searle's mystical causal powers somehow, there's

---

human memory systems are very different in that aspect, just like any other, and every person has a different mental image when thinking about the same things.

[56] This also gets rid of the need for a robot reply. Because a non-robot computer has only limited access to sensory data, in the worst case is limited to only one kind of input (which in the most cases will be text), only descriptions of said sensory data can be stored in its formal memory. That doesn't change the fact that it can have data stored, or can build data banks of stored data, concerning hamburgers.

[57] This was in chapter 3.3.1.

[58] Both the fact and the facet of a fact are just states, merely at different levels of abstraction.

nothing that makes the content in a brain specifically semantic as opposed to the content in formal memory. So I conclude that "semantic" in this context is just another buzzword, like "mental", that people like to use in order to keep believing they're special.

In the second of these posits then, Searle points out another thing: The systems reply is (at least in part) motivated by the Turing test. However, I think that's only partially correct. I think the correct way to say this would be that both the systems reply and the Turing test are motivated by the belief that there is nothing that makes us special. And it doesn't assume that computers can think,[59] rather it does assume something else: That there isn't a hidden ingredient that makes "thinking" (or mental states) special.

As to the third point, I don't think that we can reliably answer this until we've looked at what Searle means with observer-relative ascriptions of intentionality, so I'll come back to this in chapter 4.4.

[59] Of course "computers can think" is just a very blunt way of saying "computers can have states that can be called mental" – considering we found that mental states don't have anything special to them if we don't believe in mystical causal powers, in the end the belief "computers can possibly think" might well be at the basis of a belief in some sort of expressive power of the Turing test after all.

4 Intentionality in Machines

4.1 The Mind is a Machine

I shall only briefly introduce that Searle agrees with us on this particular (and very important, for the rest of this paper) subject:

> I see no reason in principle why we couldn't give a machine the capacity to understand English or Chinese, since in an important sense our bodies with our brains are precisely such machines. (Searle 1980, p. 422)[60]

I don't know if I entirely understand what he means by "in an important sense" here. I do believe that he means "except that it's not a regular machine, but rather one with added causal powers that is capable of producing mental states". All this in light of our previous and future reflections on the nature of causal powers and mental states.

So Searle essentially seems to agree that all the processes in the human brain (except the mental states, but we've seen this is pure doctrine in chapter 3.2.2) are replicable by a Turing machine, while mental states themselves are only simulatable because they lack the causal powers that would be necessary for a true replication. I'd like to look at this further though.

4.2 Replication vs. Simulation

This issue is rather important. Because Searle does seem to have a point when he compares AI to other kinds of simulations a computer can perform:

> The idea that computer simulations could be the real thing ought to have seemed suspicious in the first place because the computer isn't confined to simulating mental operations, not by any means. No one supposes that computer simulations of a five-alarm fire will burn the neighbourhood down or that a computer simulation of a rainstorm will leave us all drenched. Why on earth would anyone suppose that a computer simulation of understanding actually understood anything? (Searle 1980, p. 423)

[60] To be fair, the quote goes on like this: "But I do see very strong arguments for saying that we could not give such a thing to a machine where the operation of the machine is defined solely in terms of computational processes over formally defined elements; that is, where the operation of the machine is defined as an instantiation of a computer program." – but those strong arguments are things we're looking at elsewhere, thus my putting this part only into a footnote.

Well, this does seem convincing enough. A mere simulation of anything can't be the real thing. Searle goes on to explain why intentionality can't be created by a machine:

> Whatever else intentionality is, it is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena. (Searle 1980, p. 424)

However, you probably expected this, I don't agree. Here's why: Searle does draw parallels, however I'd like to quote Searle himself as to how far those can go and what powers parallels don't have – although he does this in an entirely different context:[61]

> But the parallel is totally irrelevant. Any valid argument whatever from true premises to true conclusions has exact formal analogues from false premises to false conclusions. Parallel to the familiar "Socrates is mortal" argument we have "Socrates is a dog. All dogs have three heads. Therefore Socrates has three heads." (Searle 1980, p. 453)

So why is this quote something that shows how the former two don't really have much explanatory or proving power? Well, it's all about parallels; Searle (just like Rorty) does something akin to a "proof by analogy" – which, as Searle rightly points out, is no proof at all. On one hand, we have the parallels between different kinds of simulations: A simulation of a thunderstorm doesn't drench, a simulation of a fire doesn't burn, so why should a simulation of intelligence intend? And, the next parallel; lactation is a biological phenomenon and dependent on biochemistry, so is photosynthesis, so why should intentionality be any different?

I do think however that Searle doesn't really go and pursue the questions he so convincingly asks long enough. Because of course there is a reason for intentionality being different from lactation, thunderstorms, photosynthesis and fires. It is implied by the following question Pylyshyn asks:

> The product of lactation is a *substance*, milk, whose essential defining properties are, naturally, physical and chemical ones […]. Is Searle then proposing that intentionality is a *substance* secreted by the brain […]? (Pylyshyn in Searle 1980, p. 442)[62]

[61] The context is: Searle is rightfully arguing against the parallel Rorty chose as basis for his otherwise quite hilarious and brilliant argumentation against Searle's theory.

[62] Emphasis Pylyshiyn's.

However, Searle dodges the core issue once again when he says (paraphrased) "no, I don't think intentionality is a substance, and I don't see what that does have to do with anything." The core issue, and the thing that makes intentionality different from lactation, is something Searle agrees upon entirely:

Intentionality is a purely logical and causal process.

This does have vast implications, and now I'm getting to the name of this very chapter. Intentionality is content, not medium. Milk and glucose (from photosynthesis) are content as well of course, they are content bound to a medium, defined by the medium. Intentionality on the other hand is abstract enough to be pure content – the medium intentionality is bound to is "mental states",[63] which in itself isn't biochemical, and maybe "common sense".[64] These mental states then are commonly realized in electromagnetical currents running through neurons.

Let's go back to analogies though, and to the simulation of a fire. A fire is a fire, but there are different kinds of fires. Blue ones, red ones, green ones – whatever the substance is that is currently burning, oxidation yields different results, has different inputs and outputs. Still, it's all called "fire". However, fire isn't replicable by a computer; it's only simulatable (unless the computer program finds a way to overheat its hardware). However, mental states aren't just simulatable, because their medium is electromagnetic waveforms[65] and their true nature is logic constructs[66] – and nothing bar intuition[67] is the thing that binds those things to the human brain.

I therefore conclude: Intentionality isn't just simulatable by a computer. It is replicable.

---

[63] We've seen already that mental states in themselves aren't special unless we assume a thing (called causal powers) that we can't define and whose sole purpose is making mental states special.

[64] Although "common sense" really only is a rationally explainable irrationality bound to those mental states, so we might just as well leave that again right away.

[65] Still, „mental states" are abstract enough not to be bound to electronic waveforms themselves. Because those electronic waveforms are just one medium on which states can thrive; and mental states are only different from other states by means of the causal powers we meanwhile exposed as myths.

[66] Just like there are different kinds of fire, nothing speaks against there being different kinds of intentionality. Maybe intentionality in a human brain is different from intentionality in a computer, just like it's going to be different from the intentionality of an intelligent alien. But still, it's all intentionality.

[67] I'm tempted to write „and common sense", in light of a previous footnote.

4.3 About Dualism

When I read Searle's paper for the first time, I was very tempted to think "now this is dualism, did Searle actually write this? He's not a dualist, right?"

Later I found that it isn't dualism at all, Searle doesn't believe that mental states are independent on their physical implementation at all – nor does he believe that the causal powers that make states mental are coming from the ether or anything, rather that the specific biochemical nature of the brain produces them.

Searle however seems to share similar beliefs when it comes to strong AI. I would like to gather some facts and relevant quotes first, before I delve into the discussion of this matter.

> Unless you believe that the mind is separable from the brain both conceptually and empirically – dualism in a strong form – you cannot hope to reproduce the mental by writing and running programs since programs must be independent of brains or any other particular form of instantiation. If mental operations consist in computational operations on formal symbols, then it follows that they have no interesting connection with the brain; the only connection would be that the brain just happens to be one of the indefinitely many types of machines capable of instantiating the program. This form of dualism is not the traditional Cartesian variety that claims there are two sorts of *substances*, but it is Cartesian in the sense that it insists that what is specifically mental about the mind has no intrinsic connection with the actual properties of the brain. (Searle 1980, p. 424)

The answer to this that Grover Maxwell gives in his commentary is quite interesting:

> Searle correctly notes that functionalism of this kind (and strong AI, in general) is a kind of dualism. But it is not a mental-physical dualism; it is a form-content dualism, one, moreover, in which the form is the thing and content doesn't matter! (Grover Maxwell in Searle 1980, p. 437)

I'm not sure I entirely agree with him though, I would have to read up on what he means with "form" and "content", and how he relates those two to the matter at hand. I'm not even sure I'd call the position that strong AI bases upon "dualism". Certainly it isn't substance dualism, so far I entirely agree with Maxwell (and Searle does, too, it seems). Searle on the other hand seems to be on the right track, particularly when he says that the actual

properties of the brain just facilitate the genesis of mental states[68] – but the mental states, being just an abstraction, a figure of thought and nothing specific or independent of their physique, can't be separated from their implementation, physically. Being just an abstraction though they can well be separated conceptually. As such, the only kind of "dualism" we might find here is a conceptual one, in that figures of thought are independent of their actual implementation.

Also, as I showed in chapter 3.3.2, I am convinced that the difference between mental states and non-mental ones is just that we like to talk about them like they're something special, while they're in fact just an abstraction; a figure of thought that makes thinking about the human brain, intentionality and intelligence (which themselves are only figures of thought), easier.

So in the end, I think that we're not talking about dualism at all here in the first place. Since mental states are just an abstraction, the real thing is what's happening in the brain. The brain produces but one particular kind of intelligence, one certain kind of mental states and intentionality.

However, just like a tire produces just one kind of elasticity (the tire-specific kind, so to say) but there's different sorts of elasticity in various vastly different materials and even structures, there's different kinds of states in all sorts and kinds of things and the only thing that keeps us from viewing them as intentional or mental is that we're not used to doing so.

4.4 Observer-Relative Ascriptions of Intentionality

What is it that defines "understanding"? Searle gives an interesting hint when he separates two kinds of intentionality:

> I think that in order to understand what is going on when people make such claims [like that a thermostat has beliefs] we need to distinguish carefully between cases of what I will call *intrinsic intentionality*, which are cases of actual mental states, and what I will call *observer-relative ascriptions of intentionality*, which are ways that people have of speaking about entities figuring in our activities but lacking intrinsic intentionality. (Searle 1980, p. 451, emphasis his)

However, how does he define the two?

> Observer-relative ascriptions of intentionality are always dependent on the intrinsic intentionality of the observers. (Searle 1980, p. 451f)

[68] Even when he does so only to show how silly that would be.

So observer-relative ascriptions of intentionality are but an abstraction, a figure of thought. Well, that's a very nice thought, and one I'd like to pursue. Because what is it that makes intrinsic intentionality intrinsic? The fact that it's based on real mental states, which in turn are caused by the mystical causal powers. Searle doesn't explicitly say that, he just states intrinsic intentionality as a fact, but he hints at it when he talks about one of the comments and says…

> Minsky says that "prescientific idea germs like 'believe'" have no place in the mind science of the future (presumably "mind" will also have no place in the "mind science" of the future). […] Even if [...] we eventually come to talk of our present beliefs as if they were on a continuum with things that are not intentional states at all, this does not alter the fact that we do have intrinsic beliefs and computers and thermostats do not. (Searle 1980, p. 452)

This seems to imply the following: the very thing that makes mental states mental, the thing that makes us different from computers and thermostats, is the thing that separates intrinsic intentionality from observer-relative ascriptions of intentionality. It is the mystical causal powers the human brain has.

I would like to argue however that intrinsic intentionality, just like observer-relative ascriptions of intentionality, is nothing but a figure of thought. This is a direct conclusion of the demystification of the causal powers – if there are no causal powers (or at least there might be special causal powers that our current AI programs don't have yet, but they're at least in theory replicable and not just simulatable), there's nothing decidedly mental about the mental, and nothing decidedly intrinsic about intrinsic intentionality.

So while it's convenient to speak of different kinds of intentionality, some of them being observer-ascribed relative intentionality and some of them being intrinsic intentionality, it's all the same in the end and there is nothing but figures of thought that, essentially, are all observer-ascribed anyway.

4.4.1 Noncognitive Subsystems, and How They Fit

Searle introduces another interesting term briefly in this section:

> The computer […] has syntax but no semantics. […] And the point is not that it lacks some second-order information about the interpretation of its first-order symbols, but rather that its first-order symbols don't have any interpretations as far as the computer is concerned. All the computer has is more symbols. (Searle 1980, p. 423)

However, that doesn't really introduce any new argumentative fuel. Because in the end, semantics are just the thing that intrinsic intentionality (or mental states, for that matter) have and observer-relative ascriptions of intentionality don't; and we pointed out already that this doesn't seem to be more than a figure of thought.

So what is it that makes noncognitive subsystems noncognitive? That we like to think about them this way, and maybe (such as in the case of Searle) that we're convinced that there must be special causal powers that mysteriously and unreplicably give true cognition, first-order symbols with meanings and intrinsic intentionality. Still, despite my disbelief in those mystical causal powers, the subject remains interesting. Because here is where the different degrees of understanding we only very briefly touched so far come into focus again.

With understanding by itself only being an abstraction, mental states only being a figure of thought and actually just being states that we like to think of as mental, noncognitive subsystems can conveniently be thought of as having mental states if we wish so.

However, we're usually aware they don't have mental states the same way a human has, and as such observer-relative ascriptions of intentionality come into play again – they're the kind of intentionality that is very limited in depth, variability, expressionality and generally complexity. That's all there is to that.[69]

4.4.2 What the Turing Test Really Tests

The general notion of what the Turing test really tests is "intelligence". But considering that this paper has come to the conclusion that intelligence, by itself being defined as "having mental states" or maybe "having intrinsic intentionality", is just a matter of figure of thought for "complex behaviour that is interesting enough so it may be described as goal- and intentionality-driven", what does the Turing test really test?

[69] Well, actually it isn't – as it never is when somebody says something like that. Searle pointing out that all kinds of non-mental systems can easily be considered to be mental under those assumptions remain, although we can legitimately refer to intuition there; because the very definitions of "mental" and "intrinsically intentional" become openly intuitive in nature. And something else entirely comes into focus all of a sudden as well: free will. However, that's a pretty large subject I'd like to not go into within this paper, just a very short version of my beliefs on this matter: Free will is an abstraction as well, in the end it's all causal interactions between parts of the "human brain machine" and its surroundings. This however shouldn't stop us from ascribing responsibility or intentionality to human beings, since those are nothing but abstractions as well – a human is responsible for his actions in the way that he is the main part of the causal system that produces his actions. As such locking away a human does make sense when that human is "wired" in a way that makes him committing criminal acts (the definition of which is a matter of convention mainly) probable.

It simply tests if various things like behavioural patterns, text recognition, fact linking and "common sensing" capabilities of any given system (or in other words, the "humane"[70] way the system handles its formal memory) are as good as a human's.[71] This might seem trivial, considering what the Turing test really is about is a human judging whether a given machine and a given (other) human are possible to tell apart in a controlled environment.

However, once we let go of the myths and intuitions that make us want to believe that the human brain is somewhat superior in reasoning power[72] to any computer, and that it is so on principle and not just because the computer isn't programmed good enough – once we let go of those intuitions the brain is just a different kind of "computer", one that happens to be biologic and not man-built.

[70] I'm using the word "humane" here due to the fact that the only real reference system we have when it comes to intelligence is the human being, we compare both animals and inanimate systems to our brain's power, and as such the main defining characteristic for a strong AI has always been how human-like it is. This is the basis of the Turing test of course, it's also the basis of much of the research that has gone into AI, and it's indeed a reasonable goal – albeit not the only one, or even necessarily the one that makes most sense. That's an entirely different discussion though.

[71] Funny enough, they can't be much better either, or the observer will find differences and definitely be able to tell which one is the computer program after all: The one that is better at analyzing complex matters.

[72] Or in the capability to produce intrinsic intentionality, of course.

5 Conclusions

5.1 States Are Just Mental If…

Searle gives us, after some initial struggle, a handhold as to what it is exactly that he believes makes states mental. It is the fact that they're both produced by and instantiated in the human brain, together with the mystical causal powers he introduces that are able to create intrinsic intentionality – which, as opposed to observer-relative ascribed intentionality, is something that only exists within distributing centres of intelligent beings, like the human brain. So what it all boils to is: The human brain has causal powers that produce "true" intentionality, while a computer doesn't.

However, once we put those causal powers under a logical microscope they reveal their true nature of only being a nice cover-up for the intuition "our brains do have got to be special", or "only humans and things we don't know yet have true intentionality, while computers and water pipes can't have that".

A very legitimate and reasonable intuition, considering most humans share it and it's one that's widely used in both literature and day-to-day conversations. However, it doesn't hold up to logical or scientific inspection, and until the day somebody finds proof that there really is such a thing as mystical "causal powers that create intentionality, who are only empirically identifiable with empiric techniques we don't know yet", it's both more convenient and more logical to assume there isn't, under the premises science currently operates.

5.2 A Circular Proof

The fact that at the very core of Searle's argumentation there's nothing but the intuition that humans[73] have to have something that a machine hasn't, and that intelligence (or intrinsic intentionality) has to be produced by some kind of causal powers whose only purpose is to create intrinsic intentionality, we're getting really circular. While Searle has this same very argument that the defendants of strong AI seem to have a circular argumentation, the thing is that Searle for his argumentation has to assume something science hasn't found to date, while the defendants of strong AI only have to assume that there isn't something science has missed so far for their proof.

As such the assumption that there has to be such a thing as strong AI isn't even necessarily important for ascribing the Turing test the message "this computer program is

[73] I'm always just talking of humans here, although animals and potential alien life forms actually fit way better into my argumentative frame than they do into Searle's – it's easier to keep the subject on just humans though, let's just keep in mind that any other life form doesn't pose any challenge whatsoever to the criterions for intelligence presented in this paper.

acting in an intelligent manner", because that just means that the computer's behaviour is similar enough to a human's behaviour so an observer can reasonably assume the same amount of observer-relative ascribed intentionality in both systems.

I doubt this can even count as circular at all, because it's not the intrinsic intentionality of the computer-program-system that's under scrutiny; intrinsic intentionality is considered a myth in the first place. In that sense, the theory is behaviourist,[74] because behaviour and intuition are the only things that make states "mental", or "intentional", considering those things are just abstractions, figures of thought with blurry edges.

5.3 Closing Words

I said in the introduction that I wanted to attempt to show how Searle's argumentation in the end only refers to intuitions in the form of mystical causal powers that somehow produce intrinsic intentionality, and I think I have succeeded in that. I also said that this wasn't meant to prove there can be such a thing as strong AI, but rather to disprove that there can't, at least not due to Searle's argumentation here. I think I have succeeded in that as well.

So while it is still possible that insurmountable problems show up that deny us the creation of strong AI, the ones that seemed to show up in the past[75] are now exposed as myths and irrational fears.

Additionally, some pointers have been given to a purely causalist point of view on intentionality and mental states, which has implications for some very important subjects like free will and human responsibility. It's a pity that those things didn't get more space in this paper, but they really were beside the point, unfortunately. Neither did ethical implications of strong AI get any space, that's a very interesting field of research as well.

Still, I hope the paper helps renew the hope of people pursuing strong AI, and helps putting the Chinese room into perspective.

---

[74] Behaviourism then becomes a way to talk about abstractions; as a matter of fact the theory presented in this paper might be a first go at attempting to resolve some pieces of the mind-body problem from a very elementary point of view. But that doesn't make psychological or behavioural efforts any less valid in that they are, unlike the theory in this paper, able to explain and predict interactions between those abstractions we're talking about.

[75] The Chinese Room was the one that seemed to be the strongest of these, so I attempted to disprove that one, I'm sure I missed plenty of other approaches that attempted to show the same though.

6 Table of Contents

7 References

Robin Gandy, 1980, *Church's Thesis and the Principles for Mechanisms*, reprinted in H.J. Barwise, H.J. Keisler and K. Kunen, eds., (1980) *The Kleene Symposium*, North-Holland Publishing Company, pp. 123-148.

Quine's Two Dogmas: Originally published in *The Philosophical Review* 60 (1951): 20-43. Reprinted in W.V.O. Quine, *From a Logical Point of View* (Harvard University Press, 1953; second, revised, edition 1961)

Scott Sehon, 2005, *Teleological Realism: Mind, Agency, and Explanation* (MIT Press)

Wikipedia, 2007: Universal Turing Machine, *http://en.wikipedia.org/wiki/Universal_turing_machine*

Wikipedia, 2007: Church-Turing Thesis, *http://en.wikipedia.org/wiki/Church-Turing_Thesis*

Wikipedia, 2007: Roger Schank, *http://en.wikipedia.org/wiki/Roger_Schank*

Wikipedia, 2007: Halting Problem, *http://en.wikipedia.org/wiki/Halting_problem*