Abstract: This thesis is concerned with the question whether the robot reply can overcome the Chinese Room argument. The Chinese Room arguments attempt to show that a computer system, executing a program, cannot have properties such as intentionality. The robot reply challenges this view, by connecting the system to the outside world, by means of sensors and mechanisms for the system to interact with its environment. If the right connections are in place, a properly programmed program can attain, for instance, intentionality. The robot reply places embodiment as the most important aspect of cognition. I attempt to show that the robot reply fails to overcome this critique. (1) The extra input given by the robot reply do not add anything. (2) A robot with the same capacities as humans does not mean it has the same experiences per se. (3) Robots are very unlike living systems.

Title: The Chinese Room and the Robot Reply | *Author*: Matthijs Piek | *Stream*: Philosophy of Science and Society | *Snr*. 2005080 | *Anr*. 872039 | *First Supervisor*: Dr. Seamus Bradley | *Second Supervisor*: Dr. Matteo Colombo

Contents

The Chinese Room and the Robot Reply

1. Introduction

Ever since the demise of the behaviorist tradition[1] in the 1950's, it became acceptable to study the internal mental states of an organism. After behaviorism, there was a cognitive turn which maintained that cognition is more or less the same as what happens in a modern digital computer (running a program). This idea is sometimes referred to as 'the computational theory of mind'. In other words: The mind is a computer program. What happens then, according to this idea, is that processes such as *thinking* are synonymous with the manipulation of symbolic representations in our brain (Thompson, 2007). In other words, if we had a properly programmed machine, running the right program, it could be said to *think*. One example of what some believed to be an indicator for, for instance, intelligence, was a machine passing the Turing test[2]. Not all computational theory of mind proponents would endorse that claim. The crux is that of running the *right* program may give a system intentionality or a mind.

The idea that a machine, which does nothing more than performing some simple syntactical operations, can have the same experiences as we do, has also been met with hostility. Our mind is more than just a computer, they claim. How can a computer suddenly have the same experiences or 'qualia' as we do? If that is the case, a calculator (with the right program installed) could, theoretically, also be said to have a mind. There seems to

---

[1] Behaviorists, most notably B. F. Skinner, claimed that we can only study the input (stimuli/conditioning) and output (behavioral response), but not the internal mechanisms (Thompson, 2007).

[2] The Turing test is a test in which a judge has to interact with a computer and a human using written questions. If the computer was successful in being indistinguishable from a human, it was argued that it was exhibiting intelligent behavior (Turing, 1950).

be something intuitively wrong with that. Is our mind just an organic calculator? The people who dismiss the claim of the computational theory of mind maintain that there needs to be more than just computations over formal symbols, perhaps some biological property. They reject that the mind is simply a device that is performing syntactical operations over formal symbols; and by extension, that a computer made on those principles can have a mind. In their view, computers (artificial intelligence *et cetera*) are nothing more than a powerful tool—something which nobody would deny.

In this thesis, I will focus on the question: Can the *robot reply* overcome the Chinese Room experiment? Some of these terms need explaining, but I will come back to them in more detail below. Simply put, the Chinese Room argument is a thought experiment that attempts to show that the mind is not merely performing calculations over syntactical elements. Furthermore, the robot reply state that it is true that only a program run on a machine with *no* causal connections to the outside world cannot have intentional agency. They argue that if we were to put the program inside a robot, with the right causal connections to the outside world, it is possible to do so.

To the main question, I will answer that the robot reply cannot overcome the Chinese Room argument. To show this, the structure of this thesis is as follows. First, I will explicate how the Chinese Room argument, put forward by Searle, shows that the computational theory of mind is a false picture of the mind. Furthermore, I will discuss the robot reply as a response to this argument in more detail. Second, I will explain the theories that lie underneath this debate, showing that the computational theory of mind is not the only game in town. Thirdly, I will explore different notions of embodiment. This is important since the way the robot is situated in its environment is, certainly by Searle, often conceptualized too simplistically. Fourthly, I will go into more detail regarding the robot reply and argue why it is regarded as a plausible response. Lastly, I will argue against the robot reply and explain a possible necessary requirement for artificial life.

One important thing to note throughout this thesis is that I refer to notions such as: intentionality, mind, consciousness, (intentional) agency, qualia, understanding, (intrinsic)

meaning and semantic content. There are obvious differences between those terms, but I am interested in whether artificial life can have any of these properties. They relate insofar that, for instance, it is thought that intentionality is a prerequisite for things like: consciousness, a mind and agency. I will use these notions to denote the possibility of artificial life possessing any of these properties.

2. The Chinese Room and the Robot Reply

To understand the Chinese Room Argument (CRA), it is essential to look at what it is supposed to demonstrate. I will briefly reconstruct Searle's (1980) thought experiment and the implications of the thought experiment as laid out in his seminal paper titled: 'Minds, Brains and Programs'. Secondly, I will discuss what the robot reply is and give a general idea of what it attempts to add. I will come back to the robot reply in more detail in Chapter 5.

I. The Chinese Room argument

Searle's CRA is targeting the computational theory of mind's (CTM) claim about strong artificial intelligence (AI), and not weak AI. What do strong and weak AI claim? Weak AI claims that AI would merely provide us with a 'powerful tool' that enables us to test a hypothesis in various fields—also referred to as 'Good old Fashioned AI'. In contrast, strong AI endorses the claim that a properly programmed computer can really have beliefs, understanding and other cognitive states. In this sense, strong AI proponents would claim that a machine, running the right program, can really be said to have a mind and that it could be used to explain human cognition (Searle, 1980). The thought that the human mind is a computational system is a position known as the CTM (Horst, 2011).

The short version of the thought experiment is as follows: Searle, who does not speak a word of Chinese, is locked up in a room. He is given boxes with Chinese symbols and a rulebook in English for manipulating the symbols. The boxes with Chinese symbols are, unbeknownst to Searle, a database and the rulebook is a program. Furthermore, he is given Chinese symbols (input) and shuffles these Chinese symbols according to the rulebook and returns these (output). Again, Searle does not know that the input are

questions and the output are answers. At one-point, Searle will become so good at shuffling the Chinese symbols, such that, from an external perspective, his answers are indistinguishable from a native Chinese speaker. His Chinese answers would be as good as if he had answered these questions in English. But, unlike his English answers, he[3] answered them in Chinese by manipulating symbols.

CTM proponents claim two things: that the program[4] *understands* the input and output and that it can help us understand human understanding. Using the CRA, we can look if these claims hold. 1) Regarding the former; Searle clearly has no understanding of Chinese, he is simply applying the rules to the Chinese symbols. 2) This shows that such programs are not sufficient to explain human understanding. Searle does not deny symbol manipulation could play no role in human understanding – after all, the program can have the same inputs and outputs as a human. Yet, there are no reasons to assume it is a necessary property of human understanding. If Searle is right, it would undermine the legitimacy of CTM, since they argue that those mental processes arise from similar mechanisms in computers by manipulating symbols. Yet, as the thought experiment has shown, it is, at most, not sufficient (Searle, 1980).

Searle (1980) states that the definition of understanding is somewhat vague, and that people have argued that there are many degrees of understanding. But he maintains that there are clear cases in which there is understanding and clear cases in which there is none. He gives the example of one's knowledge of a certain language. You might fully understand English, German perhaps less (in which case there is diminished understanding) and not understand any Chinese at all (like in this example).

---

[3] The point that *he* (Searle) is the one who answered the questions has also been contested—perhaps the 'room' answered (cf. system reply) or perhaps something else did.

[4] See footnote 3. The point is that the CTM claims that there is *someone* (or thing) that has *understanding.*

There have been many replies and objections to the CRA, but none have been conclusive (Hauser, n.d.). Some philosophers think the entire discussion is pointless. Noam Chomsky, for instance, would regard the debate about machine intelligence nonsensible. He argues that:

> "People in certain situations understand a language; my brain no more understands English than my feet take a walk. It is a great leap from commonsense intentional attributions to people, to such attributions to parts of people or to other objects. That move has been made far too easily, leading to extensive and it seems pointless debate over such alleged questions as whether machines can think" (Chomsky, 2000, pp. 113-114).

Furthermore, he observes that Turing already remarked that it is meaningless to discuss the question "Can machines think?" (Chomsky, 2000; cf. Turing, 1950). But there are those who take the CRA debate seriously.

II. The Robot Reply

The original robot reply can be rephrased as follows: If we were to put a computer inside a robot body, which is able to see, hear and act (by supplying it with sensors and mechanical arms/legs), the robot will be able to have, for instance, true understanding. To illustrate, we could give a robot a program which enables it to operate as if it: perceives, walks, eats, and so forth. The sensors and hardware in this case could be video cameras and mechanical arms and legs respectively. These sensors would enable the robot to *perceive* and its hardware to *act*, with its central processing unit functioning as its 'brain'.

Searle notes that this reply concedes that cognition cannot occur solely as "a matter of formal symbol manipulation, since the reply adds a set of causal relation with the outside world" (Searle, 1980, p. 420). The robot reply emphasizes the relation to the outside world, so this observation of Searle seems right. Therefore, if one is committed to the robot reply, it is not anymore about running *the right program*. It is also about its relation between the outside world and the robot. In this sense, the idea that a calculator can attain

understanding or intentionality[5] is false—no matter what program you supply it with. I will come back to the question of why Searle refuted this idea in a later section.

This is a shift from traditional AI approaches to a more embodied approach to AI. The shift in thinking can be thought of as going from *computational functionalism* to what we can describe as *robotic functionalism* (Ziemke, Thill, & Vernon, 2015; Harnad, 1990). One of the key ideas was that there should be a form of situatedness (Ziemke, 2001b). Situatedness is used to describe that "the physical world directly influences the behavior of the robot" (Sharkey & Ziemke, 2001, p. 253). However, what is meant with situatedness and what is needed to accomplish it has been interpreted differently. It all boils down to this question: How is a machine, or any other system for that matter, able to get semantics from purely syntactical operations? Stevan Harnad is a critic of Searle and tried to find a solution to the symbol grounding problem. The symbol grounding problem is "the problem of intrinsic meaning (or 'intentionality')" (Harnad, 1990, p. 338). Jerry Fodor, for instance, argued that a robot would derive intentional agency if the right causal connections between the robot and the environment have been established. He states:

"Given that there are the *right kinds* of causal linkages between the symbols that the device manipulates and things in the world […] it is quite unclear that intuition rejects ascribing propositional attitudes to it. All that Searle's example shows is that the kind of causal linkage he imagines—one that is in effect mediated by a man sitting in the head of a robot—is, unsurprisingly, not the right kind" (Fodor, 1980, p. 431).

This is precisely what Searle argued against as a possibility.

---

[5] Intentionality is defined by Jacob Pierre as: "[T]he power of minds to be about, to represent, or to stand for, things, properties and states of affairs" (2014). It thought to be a prerequisite for having, for instance: beliefs, desires, fears, hopes etc. It is also thought of as a pre-requisite for consciousness and a mind.

Harnad concluded that the typical reply of “symbolist” is wrong: Simply connecting a symbol system in a particular way to the real world will not cut it. Instead of this “top-down” approach of modeling cognition (the symbolist approach), he suggests a “bottom-up” approach—a hybrid response. This means that we must immerse a robot into the physical world. Only when a robot interacts with the world will it be able to ground symbols and, therefore, derive semantics. The robot is, thus, embodied and situated in the same way as humans and other animals are—he claims. Searle already anticipated such replies, which he dubbed the *robot reply* (also known as *embodied AI*) (Ziemke, 2016). In sum, embodied AI proponents claim that they can overcome Searle’s *gedankenexperiment*. Before I will talk about what it means to be embodied or situated, we will look at different cognitive theories that underpin the robot reply. It will be shown that there are non-computational options which are not attacked by Searle’s CRA.

3. Cognitive theories

In this chapter different conceptualizations underlying the robot reply will be brought to light. This robot reply can best be explained in terms of embodied AI. Specifically, there will be a focus on the underlying cognitive theories which drive embodied AI. It will be laid bare that most of the embodied AI approaches are rooted in the CTM, but that there are other ways of conceptualizing embodied AI that does not rely on the CTM. To demonstrate this, I will draw mainly on the works of Tom Ziemke and Anthony Chemero’s taxonomy.

I. A split in embodied cognition

Chemero identified two different positions within embodied cognitive science: embodied cognitive science and *radical* embodied cognitive science (Chemero, 2009). To understand the distinction between the two, it is important to look at Chemero’s taxonomy of cognitive theories. Chemero, drawing on Jerry Fodor, divides cognitive theories into two. Whilst the radical interpretation of embodied cognitive science is rooted in eliminativism, the more common embodied cognitive science theories are rooted in representationalism. I will delve into both notions more concretely, however, the spirit of

the two theories of mind are captured in the following quote by Jerry A. Fodor & Zenon W. Pylyshyn (1988).

"Representationalists hold that postulating representational (or 'intentional' or 'semantic') states is essential to the theory of cognition; according to representationalists, there are states of the mind which function to encode states of the world. Eliminativists, by contrast, think that psychological theories can be dispense with such semantic notions as representation. According to eliminativists, the appropriate vocabulary for psychological theorizing is neurological or, perhaps, behavioral, or perhaps syntactic; in any event, not a vocabulary that characterizes mental states in terms of what they represent" (As cited in Chemero, 2009, p. 17)

I will start explaining the common version of embodied cognitive science first. As stated, this version has its roots in representational theory of mind (RTM). RTM states that there are mental states, which can be, for instance, thoughts, desires, and hopes. These mental states are said to have meaning, which can be evaluated in terms of the properties it has. These intentional mental states stand in relation to mental representations. To illustrate, if we have a mental state 'The belief that the prime minister of the Netherlands is Mark Rutte', it will stand into relation with the mental representation 'Mark Rutte is the prime minster of the Netherlands' (Pitt, 2017). For instance: thoughts, beliefs, desires, perceptions will, on this view, always stand in relation with people and mental representations which represents something in the world (Chemero, 2009).

The RTM, relates to the CTM in the sense that CTM attempts to explain every "psychological states and processes in terms of mental representations" (Pitt, 2017, §8). In fact, they are very similar to each other and some use these two theories almost interchangeably (Pitt, 2017). On the CTM account, the brain is a sort of computer. Therefore, mental processes in the brain can be thought of as computations (Pitt, 2017). The projects Searle mentions in his paper are projects that rely on CTM – for instance,

"Schank's script"[6]. Embodied cognitive science is built on these conceptualizations of the mind.

Yet, this is not the only route we could take. Chemero (2009) emphasizes a "radical" approach to embodiment. This does not lead him down the path of representationalism, (which, as we have seen, tracks into CTM and traditional embodied cognitive science) but describes a route that rejects the representational theories of mind.

Chemero's second branch to embodiment is via eliminativism. Eliminativism does not hold, unlike representationalism, that the mind is a mirror of nature. They reject that the mind uses representations, and are, thus, anti-representationalists. This is something Chemero emphasizes, since his radical approach to embodiment does not rely on representationalism. It emphasizes that we can only understand cognition in relation to the life and activities of animals. Chemero defines radical embodied cognitive science "as the scientific study of perception, cognition, and action as a necessarily embodied phenomenon, using explanatory tools that do not posit mental representation" (Chemero, 2009, p. 29). Embodied cognitive science is a "watered down" version of his radical embodied cognitive science. As such, it is not targeted by Searle's CRA, since it sees cognition as purely computational.

[6] Schank Scripts can answer questions to specific stories. For instance, about a man who entered a restaurant and ordered a hamburger. The man got his hamburger, but it was burned to a crisp, and walked out angrily. The script was then asked: Did the man eat the hamburger? Which the script would answer 'No' to. Note that this was just a program executed on a computer, and there was no interaction with the environment. It was argued that 1) the machine that executed the program literally understood the story and 2) it can explain something about our own (human) understanding (Searle, 1980).

At this point, I want to attend the reader that I use this taxonomy as a general overview that underlies most of the research done in embodied cognitive science (and embodied AI). It also helps us identify which approaches to embodied AI are targeted by Searle. Yet, as Ziemke (2016) indicates, this may not give a full picture. He notes that "[Chemero's taxonomy] might not necessarily provide a complete picture, and there might be room for conceptions of cognition as a biological phenomenon that reject the traditional functionalist/computationalist view" (2016, p. 8). He expands by stating that one could "reject the traditional notion of representation" (Ibid, p. 8) which is not anti-representationalist. I am more concerned with the CTM's claim that all there is to cognition is the manipulation of *symbolic* representations or formal symbols. I am less concerned about whether mental representations (perhaps of some other kind) could play a role in cognition.

4. Embodiment

In this chapter, I will focus on the notion of embodiment, specifically, I will focus on the question of how a robot should be embodied. First, I will delve into the notion of embodiment, since it is widely held to be a necessary condition for intentional agency, yet the notion is ambiguous. Secondly, I will talk about the social aspect of the robot: Should a robot be social?

I. Different notions of Embodiment

What is embodiment? If embodiment plays a pivotal role in cognition, one might assume that there is a unified understanding of what is meant by 'embodiment'. Yet, the term embodiment is ambiguous. There are some mainstream interpretations of what embodied cognition is (Ziemke, 2001a; Wilson M. , 2002). Whilst it is true that a lot of embodied AI proponents hold it as a necessary property for cognition, it is generally unclear what is exactly meant with the notion of embodiment.

In Searle's 1980's paper, he alludes to a simple form of embodiment when discussing the robot reply (in fact, his reply is only 2 paragraphs long). Embodied, in his sense, would

merely mean attaching some sensors (e.g. auditory or sensory) and giving the robot arms and legs, for it to *act*. Then, we can put the robot in the world any way we like. That would be basically all there is to it. As long as the robot can operate in a similar fashion as humans do. Yet, little consideration has been given to *how* the body should interact with its environment (Ziemke, 2016). We can view Searle's position on embodiment as rather shallow—a simple form of embodiment that does not justify the nuances involved in embodiment.

Searle would probably maintain that it does not matter *how* you situate a robot. If the mind is all but a room where calculations are performed (Chinese room), the way you embody the system would not make any difference. What you are doing, in that case, is merely supplying the robot with more tasks for it to process. Yet, the notion of embodiment has become more important in studies of cognition.

Margaret Wilson distinguishes different claims of what is involved in embodied cognition (Wilson M. , 2002). Ziemke notes that a lot of these claims do not focus on the role of the *body* in embodiment. Whilst Wilson notes that, for instance, situatedness, time pressure, environment, and action play a pivotal role in embodied cognitive theories, it does not go into the question of *how* the body is involved into this. Only the claim that "[o]ff-line cognition is body based" explicitly does this (Wilson M. , 2002, p. 626; Ziemke, 2001a). That is why I will mainly focus on the notions of embodiment Ziemke found.

Ziemke explored the use of the notion *embodiment* too but has more focus on what the role of the body is supposed to be. Ziemke distinguishes five different notions that play a pivotal role in embodiment in the literature: structural coupling, historical embodiment,

physical embodiment and two versions of organismoid embodiment[7] (Ziemke, 2001a, p. 6). This list is structured by the narrowness of the notion of embodiment; structural coupling can be considered as the broadest, whilst the last version of organismoid embodiment as the narrowest. In what will follow, I will explain these different notions of embodiment and explicate why they are important.

First, it is claimed that to be embodied, there must be a structural coupling of the system. Structural coupling happens between two systems which continuously *perturb* each other's structure. However, it should not be perturbed in a way that is considered destructive to either system over time. One system could be, for instance, an organism (bird, tiger, bacteria, *et cetera*)—or in our case, a robot— whilst the other system could be the environment. It can be said to create a *structural fit*, whereby both systems manage to adapt to each other, which enables it to fit in their environment and *vice versa*. Both systems will come to behave in such a way, because of their intimate relationship and interaction with each other (Quick, Dautenhahn, Chrystopher, & Roberts, 1999; Maturana, 2002).

This is an important insight since it prescribes that we cannot just take an organism and fit it into a static, unmoving environment. We must also take into account the effect of the organism to the environment, and the way the environment responds to that—both systems influence each other. The consequences of this is that the system will change or evolve over time, *because* of the interplay of the two systems. Quick et al. state how this relates to the philosophical side of embodiment:

[7] Since Ziemke's publication of this paper in 2001, there have obviously been more notions of what is meant with embodiment. However, these five give a good overview of the spectrum, ranging from a broader to a stricter notion of embodiment.

"[T]here is no need to posit higher-level structures reminiscent of the folk-psychological constructs of cognitive science, such as symbols, representations and schemas. The structure at some point in time of a system coupled to its environment reflects a whole history of interaction—of the effect of environment on system, and of system on environment, which is reflected back through subsequent environmental events" (Quick, Dautenhahn, Chrystopher, & Roberts, 1999, p. § 3.1).

If Quick et al. are right that there is no necessary need to posit any symbols, representations and schemas. This would mean that the CTM is postulating superfluous conditions necessary for cognition. However, structural coupling is a very broad notion of embodiment, and creatures are embodied in a more sophisticated manner than only structural coupling.

Structural coupling is often regarded as a minimal form of embodiment. This also means that its applicability to the cognitive sciences is limited. This is mainly because it includes too much. In the sense that it would not only be applicable to cognitive systems, but also to non-cognitive systems (Ziemke, 2001a). For instance, two inanimate objects that perturb each other. More concretely, imagine that there is a rock near the shores. If the seawater continuously slams against the rock, the sea perturbs the rock. *Vice versa*, the seawater is perturbed by rock, since it changes the seawater's current flow. *A fortiori*, the fact that this can happen to inanimate objects would also mean that having an actual *body* is not a requirement for cognition. To illustrate, two (computational) systems could be programmed into a virtual environment and which mimics structural coupling. This condition is therefore clearly not sufficient for embodied cognition (Ziemke, 2001a).

Secondly, and related to the former, is the historical form of embodiment. Structural coupling cannot be merely seen as something that there is by only looking at the present. For structural coupling to have occurred, there must be a history that preceded between the systems (Ziemke, 2001a). One way to illustrate this is the theory of evolution. Over time, (on an evolutionary timescale) two systems might be perturbed in such a way that it changes its structure to better accommodate the target-environment. A species adapts itself to the environment it is currently in. But over the course of the life of a cognitive

system historical embodiment is important. The system is not just coupled to its environment in the now, but it came to be through reflection of the environment it has interacted with in the past. This does not entail physical embodiment, since this can, in principle, still be virtually simulated.

Thirdly, the system should be physically embodied. This emphasizes the fact of an actual physical robot in the environment, not merely a software program that is being executed. This is very akin to the original robot reply. Unlike Schank's scripts, there needs to be an actual physical instantiation of a robot where the program is being executed on. This also excludes the more complex virtual environments. This is a salient aspect since many researchers would concede that a requirement for cognition is interaction with the physical outside world. The best way to do so is not executing a program on a virtual machine or any other non-physical artificial environment, but to have an actual physical instantiation doing real interactions with the outside world.

Fourth, the first sense of Organismoid Embodiment (Ziemke, 2001a). This notion of embodiment stresses the fact that, for instance robots, must have certain motor sensory capabilities like other organisms. This is something also seen in the original robot reply, where a robot was supplied with sensors and mechanisms that enabled it to grasp for certain object, or to "see" and "hear". In this sense, it is more restrictive than the other notions of embodiment discussed above, since it limits embodiment to bodies that resemble organisms.

The fifth and last notion of embodiment discussed by Ziemke is also called Organismoid Embodiment but is even more restrictive. Whilst in the previous notion what was important is that the body resembles an organism, with similar motor sensory capabilities, this notion restricts it further by stating that embodiment is limited to actual living systems (Ziemke, 2001a). A living system should be autonomous and autopoietic. A system can be said to be autopoietic if it is "active, adaptive, self-maintaining and self-individuating": it enables a system to reproduce by employing strategies to regulate itself (Wilson & Foglia, 2017, p. § 2.2). An example of an autopoietic system is a cell. Cells are small factories that

produce energy, chemicals, and bodily structures from matter it extracts from its direct environment. It encompasses enzymes which perform operations on chemicals, such as snipping chemicals into two (Charlesworth & Charlesworth, 2003). It manufactures its own components and uses these components to create more of them. This is a circular process. It is autopoietic since it has a continuous self-production—also called an "autopoietic unity" (Thompson, 2007, p. 98).

Living systems are also autonomous, which means that they are self-governed. In most of the embodied AI research, for instance, there is no autopoiesis or autonomy in the system. The systems that are used in embodied AI are heteronomous and allopoietic (Ziemke, 2001a). They are heteronomous in the sense that machines are not self-governed, but other-governed. What is it to be autonomous for a living system? Thomson describes an autonomous system as follows:

"[A]n autonomous system is a self-determining system, as distinguished from a system determined from the outside, or a heteronomous system. On the one hand, a living cell, a multicellular animal, an ant colony, or a human being behaves as a coherent, self-determining unity in its interactions with its environment. An automatic bank machine, on the other hand, is determined and controlled from the outside, in the realm of human design" (Thompson, 2007, p. 37).

Living systems such as us, humans, and cells, are autonomous, because living systems do not follow the rules of others, but their own rules. As the quoted passage indicates, this is not the case for man-made machines. These machines follow the rules of 'others' and are therefore other-governed. For instance, a digital computer is given a software program which it executes. The software program consists out of commands and a rule base which describes what action it must perform and under what condition. These are programmed, not by the software itself, but by the programmers who designed it. These machines are therefore not autopoietic or autonomous, since they are not self-producing. Systems that are allopoietic are not self-producing, since they are not sustained by their own (circular) processes. These systems are called *allopoietic* (Zeleny, 1981). Digital computers are designed by humans and are allopoietic systems, which are called *heteropoietic*

(Thompson, 2007). I will come back to the differences between machines and living systems in more detail the sixth chapter.

II. Social interaction as a salient aspect of embodiment

One argument put forward is what Shaun Gallagher dubs: the *social robot reply*. Gallagher argues that we must immerse the robot into a social world. Just letting the robot out into the physical world and expecting it to put words to things is unrealistic. He suggests that we must look *beyond* embodied action in the physical environment and put more focus on *intersubjective* processes in a social world and supply it with sufficient background knowledge. Other philosophers have put forward a similar argument (Dennett, 1994; Crane, 2005).

For humans to get semantics, Gallagher argues, one must interact with "[s]peakers in physical and social contexts" (2012, p. 91). By doing so, the robot might avoid, the "framing problem". In short, the framing problem (in AI) is concerned with how a machine could revise its beliefs to approximately reflect the state of affairs in the real world (Shanahan, 2016). The framing problem is the opposite as that in humans. Whilst humans can get meaning by social interactions, relying on a vast amount of general background knowledge, robots tend to have very specialized knowledge in a specific domain. Robots do well *inside the frame* they are designed for but perform terribly outside their frame. More concretely, an AI program, designed to detect cancer cells, would still trigger positives when providing it with a picture of a car rather than an X-Ray: it is outside its framework.

To overcome this problem, Gallagher argues that we might design a robot that is able to interact and communicate with humans. This runs in multiple problems. Gallagher proposes some solutions that would enable a robot to communicate with humans. First, the robot should be able to transfer knowledge across different domains. A robot might encounter various circumstances and must be able to react differently depending on the circumstance (after all, words and gestures can have different meanings). Secondly, there

must be intersubjective interactions. A robot must meet beings that already have a lot of background knowledge. Thirdly, robots tend to be autistic in the sense that it has trouble with recognizing connections that are not direct. For instance metaphorical expressions, which often evoke associations that are not directly related to the expression itself. Most humans, however, have no problem with making those indirect associations. Therefore, a robot must find meaning wherever it can find it. Without this, it could never understand the cultural and metaphorical expressions. Lastly, it must be attuned to communication to understand the dynamics of interaction (Gallagher, 2012).

These are by no means the only ways a robot could be embodied. Ziemke (2001a), for instance, states that there could be even more restrictive by restricting it to Humanoid or human embodiment. However, these forms of embodiment will help us recognize whether they could still be a potential target of Searle's CRA.

5. The Robot Reply Continued

In this chapter, I will discuss the plausibility of the robot reply. I have already articulated the basic mechanisms of the robot reply in the second chapter, but I will go more in depth in this chapter. First, I will discuss the question of *why* the robot reply is such a plausible objection to Searle's critique. Secondly, there has been argued that when a robot has the same capacities as a human being, there would be no good reasons to argue that it does not have the same, for instance, conscious experiences as us. Thirdly, I will discuss connectionism and the embodied AI approach. Especially more recent approaches to embodied AI suggest that the right causal relations between the system and the environment have been established. In fact, they very much resemble the organismoid embodiment approach (of the first kind, still not as an actual 'living system').

I. Why the robot reply?

What makes the robot reply such an appealing idea to overcome, amongst other things, Searle's critique? The real gist of the argument lies in that the system is now receiving sensory information, which has been described by some as 'quasi-pictorial' (Harnish,

2002, p. 233). Furthermore, since there is a connection to the world, there is not just a syntactical level, but also a representational level, which is said to be semantic. It is semantic, since the representation is said to be *about* something. There are computations in the system, which stand into relation with a representation. This representation, it is argued, has content (Harnish, 2002). The reply of Searle which stated that the man in the room will be given more work, but that it has no way to know whether these symbols come from sensory input or not may therefore be wrong. After all, quasi-pictorial input may give the system more information than other input. Fodor could therefore be right, in the sense that, if the right causal connections are in place, a system may attribute meaning to those 'sensory' symbols (1980).

Another factor that made the idea of the robot reply appealing is that, in principle, we could create a robot that seem to be similarly created as other creatures. There is no reason to assume that, for instance, consciousness must be 'something' organic: So why could we not replicate it on silicon? Dennett (1994), for instance, gives the example of the robot named 'Cog'[8]. Cog was designed to be a humanoid robot. Similarly, it goes through different phases, such as infancy and it must learn itself how to use its hardware. In fact, it was designed to learn from human interaction, similarly as infants do. I will not go into full detail about Cog here (see Dennett 1994), the general point is that such a project enables a robot to 'learn' similarly as humans and attain (in principle) similar human capabilities. Why would such a robot not be able to have the same experiences as we do? All these reasons may give one the idea that there would, in fact, be nothing in the way of creating a robot that *actually* possesses intentional agency. In the next section, I will pursue the idea that a robot that has come to possess similar capabilities as human beings can be said to *have* properties such as intentional agency and consciousness.

[8] The Cog project has been stopped as of 2003.

II. The Indistinguishability of behavior output

If the robot acts as if it acts intelligent, such that humans will ascribe intentionality to it, the robot would be immune to Searle's argument. Harnad (1993) argues that if we are not able to distinguish the responses of a robot and a human (like in the Turing test) and we use these criteria in our regular life, we should not come up with new criteria to judge these robots differently. If the robot turns out to be indistinguishable from humans, this would also give us empirical support for it having genuine intentionality since we have the same capacities, Harnad argues (1993).

Jordan Zlatev agrees in this sense with Harnad and provides even further argumentation of why Harnad's claim is convincing but does raise some doubts (2001). He agrees mainly because there would be no blind symbol manipulation anymore, since the robot is wired through, amongst other things, causal connections to its environment and not "blind 'symbol manipulation'" (Zlatev, 2001, p. 160). He also notes that this provides insufficient reasons to accept that the robot possesses intrinsic meaning. However, Zlatev (2001) does give a small thought experiment which does intuitively bolster the claim of Harnad. Let us imagine that there is a person that has lived her life, similarly as other human beings. After many years the person dies. The autopsy performed on the person indicates that there is a device in her head instead of a brain. The question this raises is: Was this person a brainless machine that did not have any intrinsic meaning (Zlatev, 2001)? He states that it would be unfair to say that she did not have a mind of some sense, since it would be too late to observe the causal relations between the person's hardware and behavior. However, if we would have seen her internal hardware before she died, we would become more suspicious (depending on the implementation). To expand on this point: If we knew that the person acquired its skills the same way humans do, we might be more willing to ascribe it intentionality. Or become skeptical if all its action were pre-programmed. Thus, Zlatev agrees that it does circumvent the CRA, yet, there is still room for skepticism. However, he is still open to the possibility of such a person with genuine intentionality.

From an external perspective, we would not able to distinguish a 'person'[9] who would have no mind, with a person who does[10]. In fact, the possibility of this has been put into question (Dennett, 2013). A philosophical zombie (as Dennett calls them) is a being that is indistinguishable from other human beings, except that the philosophical zombie has no conscious experiences, intentionality *et cetera*. Dennett argues that such a zombie would, for instance, have the same odds as any other conscious humans to pass the Turing test. Furthermore, even the zombies themselves would concede that they are conscious, even though, hypothetically, this would not be the case (Dennett, 1991). In sum, there would be no way to distinguish between philosophical zombies and normal human beings.

III. Connectionism and Embodied AI

Another aspect some proponents of the robot reply invoke, is the use of connectionist models (cf. Harnad, 1990). The main goal of connectionism is to explain our intellectual abilities by using neural networks. These neural networks are claimed to be a simplification of how our brain works. They use units, which are said to be analogs to neurons that stand in relation to other units. Furthermore, the connection between different units differ in weight—some units have a strong connection, whilst other are weak (Garson, 2016).

However, traditional connectionist views have also been criticized by Searle (see next section). Furthermore, these connectionist models are often not embodied but rely on

[9] Assuming the lack of intentional agency or consciousness in the person would still constitute a person, but that is beside the point.

[10] This problem is often called: "the problem of other minds". However, only a solipsist would not attribute a mind to their peers.

artificial input and output (lacking physical embodiment) (Ziemke, 2001b). The difference between CTM and connectionism is related to representations; whilst the CTM employs symbolic representations, they are sub-symbolic for the connectionist (Thompson, 2007). Certainly, the brain-like nature of connectionism, may give more plausibility to the possibility of overcoming the CRA. Connectionist models can also easily be implemented in robots. However, I will refute this idea in the next section.

More recent approaches did take the criticism of Searle to heart and put more emphasis on the aspect of embodiment (Ziemke, 2001b). Instead of the top-down approach that traditional AI employs, it is more focused on a bottom-up approach (cf. Ziemke, 2001b). Furthermore, it does try to improve on its conception of embodiment. It is not a simple form of embodiment anymore, in which only interaction with the environment is what is important. I will address two of the principles that are used in more recent research as described by Tom Froese and Ziemke (2009).

First, it is conceded that the behavior of a system emerges from the interactions it performs with its environment (cf. structural coupling). This point was also prominent in more traditional (computational) AI approaches; however, the embodied AI approach is different. It is different in the sense that the designer of the system has less influence on the behavior of the system (Froese & Ziemke, 2009). In this sense, it has more autonomy.

Secondly is the focus on the perspective of time. There are three timescales an organism is temporarily embedded in, which Froese and Ziemke sum up as "'here and now', ontogenetic, phylogenetic" (Froese & Ziemke, 2009, p. 469). First, the here and now refers to the fact that the robot is in the (immediate) present. Secondly, the ontogenetic aspect refers to the learning and developmental stage. This relates to how an organism (or in this case a robot) develops to maturity—for instance from fertilization to adulthood in animals. Lastly phylogenetic, which refers to the evolutionary development of species. In this respect, one can see the biological aspirations. Even though it is aspired to mimic these biological features, it is not the same as living systems. For instance, evolution

(phylogenetic aspect) is artificially mimicked in a computer, which is of course, radically different that actual living creatures have evolved since they are physically embodied.

The point is that embodied AI took seriously the notion of embodiment. As we can see, some of these features recur in the stricter (but not strictest) notions of embodiment (cf. first notion of organismoid embodiment) discussed earlier. In this sense, one could argue that the approach to embodied AI very much resemble those of living organisms. If it is the case that living systems and the systems developed by embodied AI are virtually similar, it seems that the approach would be sufficient to create something that could, in principle, have intentional agency. However, in the next section, I will argue that despite the biological aspirations, embodied AI is still very much different from living systems.

6. Against the Robot Reply

In the first section of this chapter I will discuss Searle's response to the robot reply. The robot reply was already refuted by Searle in its initial paper. Searle disagreed with the robot reply proponents: "'perceptual' and 'motor' capacities adds nothing by ways of understanding […] the same thought experiment applies to the robot case" (Searle, 1980, p. 420). The man in the room will have more work, but it would not help him to attain understanding. Furthermore, I will look whether the argument made in the previous section about whether a robot with the same behavioral outputs can be said to have intentional agency.

In the second and last sections, I will discuss connectionism and embodied AI. First, I will briefly discuss whether the move to connectionist networks could in any way overcome Searle's CRA. Secondly, I will look if more sophisticated notions of embodied AI add anything. In the last section, I will discuss the shortcomings of embodied AI approaches. Specifically, there will be a focus on the difference between actual living systems and robots built on the principles of embodied AI.

I. Searle and the robot reply

Recall the CRA, where Searle was given Chinese characters. What would happen when we give the program input from the world through auditory and sensory sensors? Searle argues that it would not undermine the force of the argument. The only thing which would change is that the room is given more Chinese symbols as input and must produce more Chinese symbols as output. In fact, the man in the room would not even know that he is manipulating characters that may represent, for instance, images or sounds. In other words, all the robot does is manipulating symbols. By doing so, it cannot attach any meaning to these symbols, whilst our brain has no problem in doing so. Even if it is the case—that mental operations are done by means of computational operations over 'formal symbols'—they would have no connection with the brain (Searle, 1980, p. 424).

Furthermore, the idea that the input is 'quasi-pictorial' is highly unlikely for two reasons. First, even if we were to grant that the input is quasi-pictorial, it is not evident that the system would be able to interpret it as quasi-pictorial. It will be just more input for the system, even if given in quasi-pictorial form, since there is nothing in the 'room' that can interpret the input as pictorial. Second, the idea that input is quasi-pictorial is wrong, since bits are not arranged in a (quasi) pictorial form. The way an image or frame is built, is more complex, which also describes other properties. The system is, therefore, unable to interpret which bit string describes color, pixel-position or metadata.

Searle would therefore determine that a lot of research done in AI are a powerful tool, but that none of these (AI) systems can have true understanding. Take for example the fairly recent innovation of a robot "scientist"[11] that can do its own research (Sparkes, et al.,

[11] Note that the researchers were not interested whether the robot exhibits intentionality, but I use this example to show that AI robots are suitable as powerful tools—not to be confused with intentional subjects.

2010). The ‘Robot Scientist’ was designed to automate scientific discoveries. It does so by generating hypotheses in a domain and is able to test these hypotheses based on experiments. It can perform the experiments using robotic systems. Furthermore, it can interpret and analyze the finding based on the experimental results it has recorded. It became a news story when the robot was in fact able to make a new discovery. The robot discovered new knowledge about the genes and its relation to the metabolism in yeast. Whilst, from an external perspective, one might say that what the robot is doing is the same as a scientist does, it is not the case that the system *knows* what it is doing; let alone be capable of understanding and intentionality. This goes to show that AI should be regarded as a powerful tool, but not as something that is able to generate understanding or intentional agency (hard AI).

Would this mean that Searle thinks that consciousness is something we cannot recreate ‘something’ that simple machinery cannot accomplish? No. Denying that we could replicate, in principle, consciousness (or other mental phenomena) would propound a form of dualism. It does so by postulating *something* that cannot be studied as a physical phenomenon—whether it be *élan vital*, a soul or substance. Similarly, Searle argues that “[t]his dualism has become an obstacle in the twentieth century, because it seems to place consciousness and other mental phenomena outside the ordinary physical world and thus outside the realm of natural science” (Searle, Dennett, & Chalmers, 1997, p. 6). Therefore, Searle is not saying that it would not be possible to build a machine that can think. In fact, he states that: “The brain is a machine, a biological machine, and it can think. Therefore at least some machines can think, and for all we know it might be possible to build artificial brains that can also think” (Ibid, p. 13). The point is that merely the manipulation of syntactical elements cannot yield this on its own—as he has shown in his argument.

Searle argues that computation over formal symbols itself is not something which should be regarded as a machine process. It is differently than our brain and is very unlike the firing of, for instance, neurons or other brain processes. To contrast, computation is a process that is a mathematical and abstract process. This means that computation is observer relative, which means that the existence of the computation is relative to how

humans perceive them. This is in contrast with our brain, which is observer independent. This means that “their existence does not depend on what anybody thinks” (Ibid, p. 15).

This is an important point which Noel Sharkey and Ziemke (2001) note: The behavior of the robot only has meaning when it is attributed by an observer. In other words: The identity of the robot is contingent to an observer. To illustrate, Sharkey and Ziemke give us the famous example of Clever Hans. Clever Hans, also known under the name *Kluger Hans* in German, was a horse that was said to be able to perform actions based on additions and subtractions. For instance, if its owner would ask him the question: ‘What is three multiplied by three?’ the horse would tap his hoof nine times. Similarly, he was also able to determine which weekday it was and determine the square root of a number. Only after a German psychologist, named Oskar Pfungst, investigated it, it was discovered that Hans was only able to perform the tasks when his owner (or anyone wearing a similar hat as his) was visible to the horse. As it turned out, the owner would straighten his back or give other subtle cues when the horse had reached the right number of taps. The horse was able to pick-up on these cues and would stop tapping when the owner did so (Sharkey & Ziemke, 2001; de Waal, 2016).

This example, Sharkey and Ziemke (2001) argue, shows the analogy between the robots and Hans. Even though they might be performing arithmetic, it is not part of their natural world. Hans had no understanding of why he was tapping. Robots practically do the same. If a robot were to perform an action, and stops doing it when cued by something, it is not meaningful to the robot itself. It is, as stated, only relevant to the observer, or in Searle’s words: it is observer relative. Therefore, to attribute any anthropomorphic properties to the robot would be misguided (Sharkey & Ziemke, 2001). The only meaning that the robot has, is the one that is attributed to it.

Humans tend to attribute intentionality to other things that clearly do not have intentionality. In this sense, people will ascribe intentional attributes to objects—which

may or may-not be intentional subjects themselves.[12] To illustrate, we may use Daniel Dennett's intentional stance.[13] What the intentional stance does is making predictions based on intentional predictions. For instance, one might attribute to an object that it possesses certain information and is motivated by certain goals. Based on this, one could come up with a possible reasonable action it will 'perform' based on their held information and motivation (Dennett, 1971). One example that often accompanies this concept is a chess program. If one cannot predict the next move of the chess program by its designed response (since its design might be too complex), one can use the intentional stance. In that case, one pretends that the chess program has beliefs, mental states, and so forth. Based on those presumptions, one could possibly deduce or (inductively) infer the next rational step it may perform based on that.

Similarly, humans might ascribe intentionality to a robot. It could be the case that a certain robot has the right behavioral outputs, such that humans tend to attribute intentionality to the robot. In fact, a study found that people tended to use the 'intentional stance' towards humans and robots similarly. They stated that "the behavior enacted by the two types of agents [human and robot] were rated as similarly *intentional* and *desirable*" (Tellman, Silvervarg, & Ziemke, 2017, p. 10; emphasis in original). Yet, we must remember that this does not entail that the robot *actually* possesses intentionality, such as desires and beliefs. In this sense I stated that the only meaning a robot has is the one attributed to it.

---

[12] See for instance the (quite old) experiment by Fritz Heider and Marianne Simmel, in which they showed an animation, consisting of basic shapes. It was shown that, even though it concerns basic shapes, people interpret the actions of those shapes as actions of persons (Heider & Simmel, 1944).

[13] It might be helpful to point out that an intentional system does not have to be an intentional agent *per se.* As Dennett (1971) points out "The concept of an Intentional system is a relatively uncluttered and unmetaphysical notion, abstracted as it is from questions of the composition, constitution, consciousness, morality, or divinity of the entities falling under it" (p. 100). Therefore, and Dennett (1971) mentions this: It is easier to conceive of something as an intentional system than it is to say that a machine can really think.

There are further grounds to state that the robot cannot have intrinsic meaning if it has the same capacities as us. After all, if a robot, which has the same behavioral outputs as humans do, is said to have no intentional agency by the mere fact that it is a robot, it would totally exclude the possibility of robots as intentional agents. Recall Harnad's claim that if we cannot distinguish the capacities of a robot from a human (for instance Dennett's philosophical zombies), we have no reasons to postulate new criteria, since we use the same criteria to determine if other humans possess intentions. Should we therefore concede that in these cases robots do, in fact, possess intentional agency? Yes and no. Regarding the former, I concede that it would be impossible to discern "philosophical zombies" (insofar as they can exist) from other people, and do not intend to attempt solve that problem here. In that sense, it could be the case that a robot does in fact have intentional agency. Regarding the latter, which Zlatev already stated, we should become more skeptical if the implementation is pre-programmed or that it acquired knowledge of this world differently than other creatures like us. He states that:

"[T]he pre-programmed robot will have nothing but (first-order) *derived intentionality* – all its 'representations', 'goals', 'beliefs', etc. would derive their meaning entirely from the intentionality of the engineers who programmed it" (Zlatev 2001, 162; emphasis in original).

Therefore, if we found out that the implementation is such that no intrinsic meaning could be formed, the only meaning that there is, is the one attributed to such a robot. To connect this to the critique of Searle: If the robot is still based on the CTM (thus susceptible to the CRA), then it would mean that, even if the robot has the same capacities as human beings, cannot attain intentional agency. Or in Searle's words: "[A] system could have input and output capabilities that duplicated those of a native Chinese speaker and still not understand Chinese, regardless of how it was programmed" (1980, p. 423).

II. Connectionism

Searle (1990a) has also responded to the question whether systems based on connectionist models would overcome his argument. He claims, at least for traditional

connectionist models, it does not. Searle does seem a little bit sympathetic in his view towards connectionist models vis-à-vis traditional computational systems. He states that:

"[A]t least some connectionist models show how a system might convert a meaningful input into a meaningful output without any rules, principles, inferences, or other sort of meaningful phenomena in between. This is not to say that existing connectionist models are correct – perhaps they are all wrong" (Searle, 1990b)

Searle does seem to refute the idea. He states that: "The parallel, 'brainlike' character of the processing […] is irrelevant to the purely computational aspects of the process" (Searle, 1990a). To show this, Searle came up with a tweaked version of his CRA, which he dubbed 'Chinese gym'. The basic mechanisms are like the original CRA, so I will briefly recapitulate the thought experiment and explain what it attempts to show.

Searle envisions a hall containing multiple monolingual men. These men perform similarly to the connectionist models, in which the men represent the different nodes and synapses. The outcome, Searle claims, would be the same as if one man were to perform the operations. No-one in the hall can speak a word of Chinese, and there would be no way for them to do so (Searle, 1990a).

If the connectionist network is, thus, based on purely computational principles (akin to claims by the CTM), it would still have the same problem as traditional AI systems have. Even though it may give a more accurate picture of brain-like-activities, it is not able to escape Searle's Chinese room. Furthermore, connectionist approaches are more concerned with mimicking the brain, there is often no concern for sensory and motor coupling between the 'brain' and the environment (Thompson, 2007). Since we are concerned with the robot reply, we could put such a connectionist system in a robot body to see if anything changes. However, this would not change the force of Searle's argument. It would run into the same trouble as traditional CTM has with Searle. There is a catch, since Searle is concerned with the CTM, there are connectionist systems that are

not targeted by Searle's CRA (in this case Chinese Gym). For instance, if it is like Chemero's (2009) radical form of embodied cognition.

III. Organisms versus Man-made Machines

I will now look at embodied AI and inquire into whether it could, in principle, be enough for intentional agency. Embodied AI inspires to a form of organismoid embodiment, but, as Ziemke (2016) points out, there is still a fundamental difference between living systems and robots.

Ziemke (2016) notes that much of embodied AI research seems to ignore theories such as autopoiesis. He states that researchers treat the robot body as a simple input and output device, which, like Harnad, must generate physical grounding. This means that whilst the embodied AI researchers try to attain something similar as biological embodiment in AI, they fail to do so. Ziemke expands by saying that "there is no integration, communication or mutual influence of any kind between parts of the body, except for their purely mechanical interaction" (Ziemke, 2001a, p. 9). This means that there is still a dichotomy between mind and body, or software and hardware. The physical body is mainly seen as something that serves the 'mind' to interact with the physical environment (Ziemke, 2016). If we look at organisms, however, this dichotomy between mind and body is blurred. The organism's body is not passively integrated with the mind.

Albeit researchers aspire to mimic living systems, they fail to do so. As we have seen in the previous section, embodied AI tries to incorporate aspects such as 'the here and now', but also consider the ontogenetic and phylogenetic development of their robots.
Ziemke underlines the difference between living systems and artificial robot bodies: "[I]n the evolution of robot bodies […] there is no growth or adaptation of the individual robot body. Instead body plans are first evolved in the computer, i.e. 'outside' the robot, and then implemented in a robot body" (Ziemke, 2001b, p. 219). If this is the case, then the possibility of self-organization and autonomy is undermined (or, at least, diminished) since they did not construct themselves, but according to a plan made by the engineer (Ziemke,

2001b). Furthermore, the intersubjective aspects, which Gallagher discussed, is often ignored.

Another difference is that in embodied AI the robots have no intrinsic stake in the world. The engineers put in place goals which the system is directed towards. These goals are externally imposed by the engineer and there is no good reason to assume that those goals would become intrinsic to the system. To illustrate, Froese and Ziemke (2009) give the example of a robot that has specified motivational states. The motivational states are based on inputs which encode hunger and thirst. Talking about the goal or the agenda of such systems is, therefore, purely metaphorical, since the defined goals are not goals the system has defined for itself. When it may seem that the robot desires to drink as to avoid becoming thirsty, it has no intrinsic meaning to the robot. The only meaning that can be attached to this 'desiring', is external (for instance, by the engineer) (Froese & Ziemke, 2009).

The point is that robotics, even though they embody their robotic systems, is still very much different than actual *living* organisms. After all, the mere fact that a robot has a battery pack, which it must recharge at times, is allopoietic and certainly not self-maintaining. One might raise the objection that we could, for instance, supply the robot with solar panels to create a more self-maintaining robot. Even if in some sense self-maintaining, some fundamental differences persist. The robot can be shut down when its batteries run-out. When recharged, there is not necessarily anything that is lost when the robot is turned back on again. In contrast, organisms cannot be totally shut down and restarted like robots. In other words: Whilst robots can be said to be 'reversible', living systems are 'irreversible' (Bickhard, 2009).

It seems that Searle (1980) is right in stating that intentionality is not purely computational. If we want to create 'strong AI', there must be some biological aspect. If it is true that

autopoiesis is a necessary[14] condition for cognition and that robots are allopoietic, it would follow that robots that are not embodied in the strictest sense (second form of Organismoid embodiment) cannot have a mind. Autopoiesis as a necessary condition for intentionality would also concede Searle's point that intentionality is a biological phenomenon. Searle regards intentionality as "causally dependent on the specific biochemistry of its origins" (Searle, 1980, p. 424). Autopoiesis could be this biological phenomenon that would enable a system to truly have intentional agency. If we want to create these systems, we must regard cognition as a biological phenomenon. Simply running the *right* program (surrounded by allopoietic hardware) cannot do this by itself.

If Thompson is right that autopoiesis is a necessary condition for cognition, it would mean that current hardware must be replaced with 'autopoietic hardware', which, as far as I know, has not been forged. There could, of course, be other (biological) phenomena (cf. Ziemke, 2016; Froese & Ziemke, 2009) that are required for intentionality. However, by establishing that biological phenomena play a role in creating a 'mind' or intentional agency, the claim of CTM that the mind is purely computational is false. In this sense, whilst some embodied AI approaches use the less restrictive notion of Organismoid embodiment, this conclusion pushes us in the direction of the strictest notion of embodiment.

We must remember that the human mind is very far from being understood. The fact that, almost four decades after Searle's publication, there is still research and papers being written on the topic. In fact, over 2017, Searle's paper (Minds, Brains, and Programs) has

[14] It has also been argued (controversially) by Thompson that autopoiesis could be a sufficient condition of cognition. Thompson states that: "[I]f autopoiesis is taken more widely to mean internal self-production sufficient for constructive and interactive processes in relation to the environment, then autopoiesis does entail cognition" (2007, p. 127).

been cited over 200 times. This could be since the cognitive sciences are a relatively immature field—something which Chemero (2009) has indicated. To exemplify this, we may look at the myriad of purely logical thought experiments (e.g. Searle's CRA) are used in the cognitive sciences. These are claims that could, in principle, be established empirically (Chemero, 2009). However, as I have tried to show, there are good reasons to regard cognition as a biological phenomenon.

7. Conclusion

In this thesis, I have looked at the Chinese Room argument by Searle and whether the robot reply could overcome this argument. I have shown that it does not. I have identified different cognitive theories that underlie the robot reply, which showed that most of embodied AI are based on computationalist theories. However, there are also conceptualizations in which this is not the case (cf. radical embodied cognition). Furthermore, I have explored the notion of embodiment, which showed that it is ambiguous. There are different notions, some which can be regarded as very broad (e.g. structural coupling) and more narrow notions, for instance, organismoid embodiment. The robot reply does try to create interaction with the environment that is like other organisms (e.g. ontogenesis, phylogenies). Some have argued that the robot reply is able to overcome Searle's thought experiment since it provides causal connections to the outside world. This means that, if the right causal connections are in place, a robot could have intentional agency. Furthermore, if we can create machines that have the same capacities as us, and are grounded in those purely computational principles, we would have no way to know if they have the same experiences as us (and have to assume they do).

Searle noted that any system sees the mind as solely manipulating formal symbols, cannot escape the Chinese room. Artificially mimicking things, like ontogenesis, in those cases, would not help a bit. Even if the robot were to have the same capacities as a human being, and we knew how it was programmed, we could become more skeptical and refute the idea that such a robot could have the same experiences as us. Furthermore, even though robotics aspires to a biological approach, it fails to do so—robots are very different from living systems (cf. organismoid embodiment/social aspect). Could computations play

a role in cognition? Perhaps, but I think Searle has convincingly shown that it is not a *sufficient* explanation for things like intentional agency. Therefore, I think that Searle is right in stating that a computer (or robot) cannot, for instance, possess intentional agency or be conscious solely by executing a program which manipulates formal symbols. Furthermore, Searle has shown that the CTM is false and putting a computer in a robot body would not overcome the CRA.

8. Bibliography

Bickhard, M. H. (2009). The biological foundations of cognitive science. *New Ideas in Psychology, 27*(1), 75-84.

Charlesworth, B., & Charlesworth, D. (2003). *Evolution: A Very Short Introduction.* New York: Oxford University Press.

Chemero, A. (2009). *Radical Embodied Cognitive Science.* Cambridge, MA: MIT Press.

Chomsky, N. (2000). *New Horizons in the Study of Language and Mind.* Cambridge: Cambridge university press.

Crane, T. (2005). *The mechanical mind : a philosophical introduction to minds, machines and mental representation.* Routledge. Retrieved from http://www.routledge.com/books/details/9780415290319/

de Waal, F. (2016). *Are We Smart Enough to Know How Smart Animals Are.* London: Granta Publications.

Dennett, D. (1971). Intentional Systems. *The Journal of Philosophy, 68*(4), 87-106. Retrieved from http://www.jstor.org/stable/2025382

Dennett, D. (1991). *Het bewustzijn verklaard.* Amsterdam: Olympus.

Dennett, D. (1994). *Consciousness in Human and Robot Minds*. Retrieved from http://people.ku.edu/~mvitevit/dennett_humanrobotminds.pdf

Dennett, D. (2013). *Gereedschapskist voor het denken.* WW Norton & Company.

Fodor, J. A. (1980). Searle on what only brains can do. *The Behavioral and Brain Sciences*, 431-432.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*(1), 3-71.

Froese, T., & Ziemke, T. (2009). Enactive artificial intelligence: investigating the systemic organization of life and mind. *Artificial Intelligence, 173*, 466-500.

Gallagher, S. (2012). Social Cognition, the Chinese Room, and the Robot Replies. In R. Zdravko, *knowing without thinking* (pp. 83-96). Palgrave MacMillan.

Garson, J. (2016). Connectionism. *The Stanford Encyclopedia of Philosophy*.

Harnad, S. (1990). The Symbol Grounding Problem. *Physica D, 42*, 335–346.

Harnad, S. (1993). Grounding symbols in the analog world with neural nets. *Think, 2*(1), 12-78.

Harnish, R. M. (2002). Criticisms of the Digitial Computational Theory of Mind. In R. M. Harnish, *Minds, Brains, Computers* (pp. 225-270). Massachusetts: Blackwell Publishing.

Hauser, L. (n.d.). *Chinese Room Argument*. Retrieved 11 1, 2017, from http://www.iep.utm.edu/chineser/

Heider, F., & Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology, 57*(2), 243-259.

Horst, S. (2011). Horst, Steven. "The computational theory of mind. *Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/entries/computational-mind/

Maturana, H. (2002). Autopoiesis, Structural Coupling and Cognition. *Cybernetics & Human Knowing, 9*(3-4), 5-34.

Pierre, J. (2014). Intentionality. *The Stanford Encyclopedia of Philosophy*.

Pitt, D. (2017). Mental Representation. (E. N. Zalta, Ed.) *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/spr2017/entries/mental-representation

Quick, T., Dautenhahn, K., Chrystopher, N. L., & Roberts, G. (1999). The Essence of Embodiment: A Framework for Understanding and Exploiting Structural Coupling Between System and Environment. *Third International Conference on Computing Anticipatory Systems*, 649-660.

Searle, J. (1980). Minds, Brains and Programs. *The Behavioral and Brain Sciences*, 417–424.

Searle, J. (1990a). Is the Brain's Mind a Computer Program? *Scientific American*, 26-31.

Searle, J. (1990b). Consciousness, explanatory inversion, and cognitive science. *Behavioral and Brain Sciences*, 5885-642.

Searle, J., Dennett, D., & Chalmers, D. (1997). *The Mystery of Consciousness.* New York Review of Books.

Shanahan, M. (2016). The Frame Problem. (E. N. Zalta, Ed.) *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/spr2016/entries/frame-problem/

Sharkey, N. E., & Ziemke, T. (2001). Mechanistic versus phenomenal embodiment: Can robot embodiment lead to strong AI? (R. Sun, Ed.) *Journal of Cognitive Systems Research, 2*, 251-262.

Sparkes, A., Aubrey, W., Byrne, E., Clare, A., Khan, M. N., Liakata, m., . . . King, R. D. (2010). Towards Robot Scientists for autonomous scientific discovery. *Automated Experimentation, 2*, 1-11. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2813846/

Stuart, J. A., & Noel, S. E. (1996). Grounding computational engines. *Integration of natural language and vision processing. Netherlands: Springer*, 167-184.

Tellman, S., Silvervarg, A., & Ziemke, T. (2017). Folk-Psychological Interpretation of Human vs. Humanoid Robot Behavior: Exploring the Intentional Stance toward Robots. *8*, 1962. doi:10.3389/fpsyg.2017.01962

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind.* Cambridge and London: Harvard University Press.

Turing, A. M. (1950). Computing Machinery and Intelligence A.M. Turing. *Mind, LIX*(238).

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9*(4), 625–636. Retrieved from http://www.indiana.edu/~cogdev/labwork/WilsonSixViewsofEmbodiedCog.pdf

Wilson, R. A., & Foglia, L. (2017). Embodied Cognition. (E. N. Zalta, Ed.) *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=embodied-cognition

Zeleny, M. (1981). What is autopoiesis." Autopoiesis: a theory of living organization. *New York: Elsevier*.

Ziemke, T. (1999). Rethinking Grounding. In A. Riegler, M. Peschl, & A. von Stein, *Understanding Representation in the Cognitive Sciences* (pp. 177-190). Boston, MA: Springer.

Ziemke, T. (2001a). Are Robots Embodied? *First international workshop on epigenetic robotics Modeling Cognitive Development in Robotic Systems (Vol. 85)*, 4 - 11.

Ziemke, T. (2001b). The construction of 'reality' in the robot: constructivist perspectives on situated artificial intelligence and adaptive robotics. *Foundations of Science, 6*(1), 163-233.

Ziemke, T. (2016). The body of knowledge: on the role of the living body in grounding embodied cognition. *Biosystems, 148*, 4-11.

Ziemke, T., Thill, S., & Vernon, D. (2015). Embodiment is a Double-Edged Sword in Human-Robot Interaction: Ascribed vs. Intrinsic Intentionality. *Cognition: a bridge between robotics and interaction. Workshop at HRI2015*, 1-2.

Zlatev, J. (2001). The Epigenesis of Meaning in Human Beings, and Possibly in Robots. *Minds and Machines, 11*(2), 155-195.

Zlatev, J. (2003). Meaning = Life (+ Culture): An outline of a unified biocultural theory of meaning. *Evolution of Communication*, 253-296.