# Philosophy and Ethics C&IS
# Final assignment

Menno Bartels s1007797
Lucas van der Laan s1047485
Marco Post s1046670
Elwin Tamminga s1013846
Ernst Hamer s1047547

January 28, 2022

## 1 Introduction

In recent years, artificial intelligence (AI), is mainly used for solving problems, that are hard for humans to solve. An example of a current useful application, where AI surpasses humans is folding proteins. Folding a protein is the process of determining the three-dimensional structure of a protein, given its protein chain. This is a task that a human can perform, with great effort and attention, resulting in a reasonable outcome. However, we humans do not fully understand this process yet. A lot of work involved in this research area is trial and error, or so-called, experimental research [4]. The AI system that is on its way to solving this problem is called AlphaFold [12]. The system tries to achieve the same task as a human would do, namely folding the protein in the correct way. Only, this system shows results that are more impressive than the results of human scientists. "The model from group 427 gave us our structure in half an hour, after we had spent a decade trying everything," Lupas says [2]. Quite impressive. We can say that the AI has learned how to fold proteins, or, in other words, has learned a more efficient way to do it than humans. It could be said, that to some degree, the system is more intelligent than humans. Such new developments in AI have led to interesting discussions.

If there is no fundamental difference between human intelligence and artificial intelligence, then we need to take this into account when developing technologies that can imitate human behaviour using artificial intelligence. An AI that not only behaves like a human but also has a mind like a human, is considered a strong AI. John Searle is a philosopher who presented arguments against such an AI. In this paper, we discuss these arguments and use Searle's definition of a strong AI: "The appropriately programmed computer with the right inputs and outputs would thereby have a mind in exactly the same sense human beings

have minds" [10]. We will also classify strong AI in Aristotle's classification hierarchy of souls (figure 1).

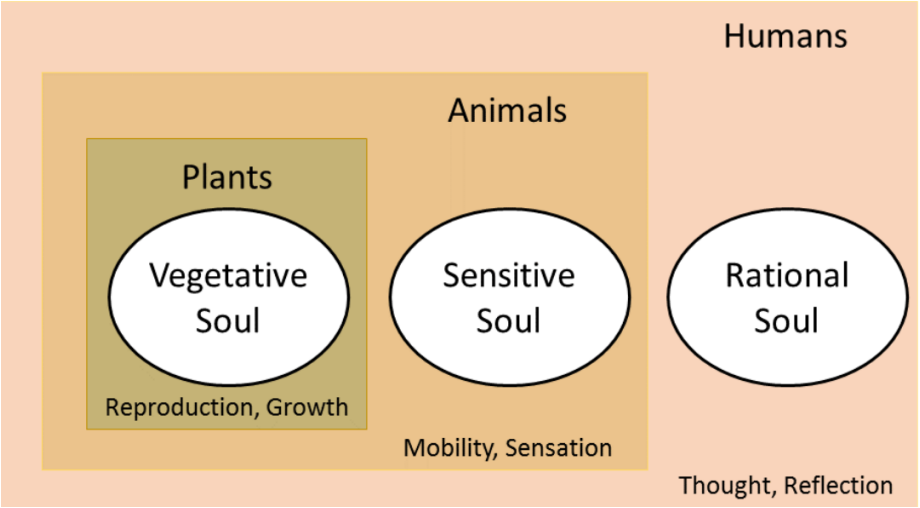


Figure 1: Aristotle 384-322 BCE (Before Common Era), Man as rational soul

Our research question, comes forth from these discussions on recent advancements in AI systems. Although these recent advancements in AI, similar to Alphafold and GPT-3, are not yet on the level of strong AI. We speculate that eventually, research in AI systems is to take on the likes of strong AI described above.

The research question we addresses is the following: "Where to classify machine (strong AI) in Aristotle's classification hierarchy of souls?"

In this paper, we argue that strong AI falls into the same classification as humans. This is something we will defend using arguments against the Chinese room argument that Searle gives [9]. We will explain why we think arguments in favor of the Chinese room are invalid, by taking a stance in Ludwig's Inconsistent Tetrad.

We will discuss the Tetrad, and explain which one of the standpoints we agree with. With this standpoint we will conclude with the explanation of why strong AI is classified the same way as humans in Aristotle's classification hierarchy of souls.

# 2 Body

In 1980 John Searle used the Chinese room experiment to show that the Turing test is insufficient to show that a system can be described as having strong AI. A strong AI is a system that actually understands what it's saying, and a weak AI is a system that emulates understanding. In the Chinese room experiment, a box emulates the understanding of Chinese text, without really understanding. Using the Chinese room, Searle argues that computers programmed to show intelligence, do not have intelligence.

How Searle described a machine might simulate intelligence is not compatible with how modern AI researchers think human-level intelligence in machines will be achieved. The Chinese room receives information and outputs a response without gaining feedback on what response it gave. Modern AI systems, in general, take in information and try to update a model based on some feedback it's getting from the outside world. The updating of the model is performed such that the function is generalizable over all possible inputs for that specific task. This process, called machine learning, creates a model that is highly proficient in a single task. This process of training is highly generalizable over many differing tasks. It is still not clear that machine learning will achieve strong AI someday. AI researchers have disagreements about how and when human-level intelligence will be achieved, but the general belief is that human-level intelligence in machines will be reached someday [3].

To classify machines in Aristotle's classification hierarchy of souls, we can use the arguments against, and in favour of the Chinese room experiment. We relate these arguments to Ludwig's Inconsistent Tetrad [7].

## 2.1 Realism

The first point of the Tetrad seems obviously true. It states that the Mental exists and some things have mental properties. In this section 1.2 of [7], "the Mental" is defined as having consciousness. Then, having consciousness is defined as being able to have conscious mental states, which we can discriminate, remember or forget. Conscious mental states include perceptual experiences, somatic sensations, pains and itches, feeling sad or angry, or hunger or thirst, and occurrent thoughts and desires. This can be put into other words. Namely, that being conscious is defined as being able to perceive what is happening, and to be able to distinguish, remember and forget events that one has perceived in the past. Now we can prove that the Mental exist with the following. Because of the fact that you are reading this text, and thus perceiving what is. If the mental didn't exist, the reader wouldn't be able to read and understand the meaning of this text just by looking at it. The first point of the Tetrad then also claims that some things have mental properties. Again we use the same argument as we used for proving that the Mental exists. You are reading the

text right now, because of that, you have the mental property of perceptual experiences. It does not matter if you are a human or an AI, it only matters that there is something reading this text. The thing, that has the mental properties.

## 2.2 Conceptual autonomy

The second point of Ludwig's Inconsistent Tetrad is the point that mental properties are not reducible to non-mental properties. Which can be recognized as the simple formula; "mental phenomena $\neq$ physical phenomena". This is the point of the Tetrad we reject in our paper on the basis of the arguments stated below.

The first argument we make for our position is that, if the mental can be reduced to the non-mental, then, conceptually, non-mental propositions must entail mental propositions. This small step to take seems logical, however, there are some problems with this conceptualization. For example, we humans can imagine non-material, mental things, that could range from ghosts to gods, which shows us directly that for the mental, physical material is not conceptually necessary. In our second argument, we turn the first given argument on its head. We can make the first argument the other way around, such that we can imagine material, non-mental things, thus presenting us with the information that non-mental propositions do not necessarily entail mental properties. Examples following this line of argumentation are biological robots mimicking (parts) of human behaviour. In this example the physical material of the biological robot has no real mentality or consciousness, thus the material is not conceptually sufficient for the mental. These two points, arguing from both sides of the equation of conceptual autonomy, present us with a contradiction at the core of this idea. Thus, it does not necessarily seem that the mental can be reduced to the non-mental.

If mental cannot be reduced to non-mental, then an AI would never be able to think like a human in the Chinese room experiment because of the missing mental capabilities of a human, even if it would get feedback. Thus by accepting conceptional autonomy automatically implies that the Chinese room argument is correct. We think that Searle does not reject conceptual autonomy though, he thinks that a machine cannot think because it does not have a "mind" [9]. But by rejecting conceptual autonomy, non-mental propositions (e.g. computations) can entail mental propositions, which means that machines can have a "mind". Some philosophers refute Searle's argument because of this reason [6], because it would invalidate the Chinese room experiment itself. Searle himself doesn't think he is a property dualist, because he says "consciousness is a mental and therefore biological and therefore physical feature of the brain", which according to him not a property of dualism [11]. This could imply that he does not really accept conceptual autonomy, but some philosophers do not agree that he is not a dualist [5]. We argue that if Searle really believes that consciousness is a physical state in a biological and physical brain, he must then contend

with the fact that non-biological physical systems have been created that do not fit the intelligence simulation conclusions from the Chinese room experiment. Namely, Searle didn't allow for the Chinese room to learn from feedback, which modern AI systems do. This means that, because modern AI systems can learn, the Chinese room experiment is no argument against the possibility of strong AI.

## 2.3 Constituent explanatory sufficiency

The third point in the Tetrad states that constituents of a whole can sufficiently explain its behaviour, i.e., everything can be described as the sum of its parts. If one would reject this claim, one would say that something is more than the sum of its parts. This philosophy is called emergentism and is then contrasted with reductionism. While emergentism and reductionism are opposites, they both strive to conform with physicalism [13]. In other words, the mental is still fully emergent from only physical entities and every physical event is fully accountable to only physical causes. This leaves no room for mental causation i.e. the mental influencing the physical. Some emergentists argue that the mental is not fully reducible to 'simple' materials. From certain organizations of these simple materials, complexity can emerge. The mental is one of these complex emergent properties. One explanation of this was made by Huxley (Darwin's bulldog). He argued that consciousness was "a shadow cast by neural activity" [7]. Mental activity was not the goal of neural activity, but merely a consequence. This argument is also called epiphenomenalism, which holds that mental properties are not causally relevant to anything. However this is self-contradictory: if we have knowledge about epiphenomenalism, then our brains knows about the existence of the mind, but if epiphenomenalism were correct, then our brains should not have any knowledge about the mind, because the mind does not affect anything physical. If emergentism were to be true, epiphenomenalism must also be true, which refutes the physicalist principle of causal closure.

Searle's argument that strong AI can't emerge from computers programmed to simulate intelligence is an argument for constituent explanatory sufficiency. Because computers can only be programmed to emulate understanding, but don't actually understand, and understanding will never arise in programmed machines, intelligence will never emerge. Because Searle is making a reductionistic and physicalistic argument here, it leaves room for downward causations, i.e, higher-order systems (the mental) can cause changes in lower-order systems (the brain). Downward causation however can, and is, programmed into computers. Supervised learning is an example of downward causation. In supervised learning a lower order system, the model, gets feedback from the outside world, a higher-order system, to update itself using some the feedback and some update function.

Using the definition of emergentism and downward causation we can conclude Searle was neither an emergentist nor did he account for downward causation

in the Chinese room experiment. Consequently, we can conclude that although strong AI is not emergent this is not a valid argument for the Chinese room experiment conclusions.

## 2.4 Constituent non-mentalism

The fourth point states that "The basic constituents of things do not have mental properties as such" [7]. This is supported by the fact that the body of a human exist of a lot of basic constituents, like flesh and blood, that do not have a mental capacity of their own, instead the mental in the human body is a combination of neurons in the brain that steer the entire body's functionality. The basic constituents themselves do not have mental capabilities, but together, they create the mental capability [8]. So for an AI, this is the same, as an AI is just composed of "neurons" in a neural network, which imitates how a human's brain works, so to deny that an AI has mental properties based on this would be denying that a human has mental properties.

If we believe that a human indeed only has mental properties with all the neurons together, instead of the mental being a non-physical property related to something like a soul, then we can also state that an AI with enough neurons and enough nurturing, just like a human child, could indeed create mental properties. This would then indicate that a strong AI should be capable of learning, thus going in opposition to Searle's Chinese room proposition. A human baby needs to learn from its surroundings using all of its senses that have been adapted for this fact through natural evolution, an AI does not have this naturally, because humans create an AI and thus would have to know enough about the brain of a human before they can then create a strong AI capable of copying or exceeding human intelligence.

If we were to instead believe in something like a soul that creates mental properties instead of the physical neurons in the brain, then we could never agree to the fact that a strong AI can exist, as it bears no soul to carry its mental properties, like thought and understanding what it is doing.

There is also Panpsychism [1] which deems that everything in the universe has mental properties. This means that things like trees, stones, and even molecules have mental capabilities and thus live. If you were to believe in this, then you would reject 4, as a human may have mental properties, but so do the components of a human body. This would also mean that a computer would have to be conscious, which means that strong AI would by definition exist, as it is sentient and can think. If it would pass the Turing test is a different story, as that would mean that the AI has indeed learned, which would mean that either the AI has figured out how humans work on its own, or a human has taught the AI how to learn to be a human.

### 2.5 Reject 2 - Conceptual autonomy

Ludwig's Inconsistent Tetrad consists of four propositions of which three can be true, and if one is rejected, the others are accepted. Of the Tetrad points mentioned above, we accept the points on Realism, Constituent explanatory sufficiency, and Constituent non-mentalism. This means that we reject point 2, on Conceptual autonomy. Rejecting the point on conceptual autonomy means in this case that we deny that mental properties cannot be reduced to physical properties. This can lead to the following viewpoints relating to strong AI: Logical behaviourism, Identity theory, and Functionalism. As we have argued in the body of this paper, rejecting this point of the Tetrad, means we reject the conclusions John Searle made from the Chinese room experiment. Thus, providing justification for classifying strong AI, in Aristotle's hierarchy of souls, on the matched position of man as rational soul, with the ability of at least human-level thought and reflection.

## 3 Conclusion

The research question of this paper is, "where to classify machine (strong AI) in Aristotle's classification hierarchy of souls?". In our analysis, we have worked out the four points of Ludwig's Inconsistent Tetrad, and in doing so, explained which points we accept, and reject. On the basis of rejecting the second point, we argued that we reject the Chinese room experiment as an argument against the possibility of strong AI. This places strong AI at the same ranking position as humans, in the hierarchy of souls of Aristotle.

Our conclusion is based on rejecting point two of Ludwig's Inconsistent Tetrad because we think that mental properties can be reduced to physical properties. However, we don't know for sure if this is true, because it is still a theory, till it is proven that souls actually exists, which we think is a long shot. This means we can only assume that Searle's argument is false until it's proven or disproven that mental properties can be reduced to physical properties. This also depends on the definition of a mind. For example, if the mind of an AI behaves exactly like a human mind but cannot be considered the same because of the physical properties it consists of. However, by our definition of a mind and rejecting the second point of the Tetrad, we do think that a strong AI can be classified as a rational soul in Aristotle's classification hierarchy of souls.

As a concluding remark, we would also like to express that, one could doubt the whole conceptualization on which the hierarchy of souls is based on, and thus the big picture of our paper. Although Aristotle was one of the sharpest minds of his time, his time was still 384-322 BCE, where they perhaps had different ideas about rational souls, human thought, animals thought, and reflection. The biological ideas of minds in animals have quite progressed since then, with some biologists arguing that some animals (e.g. dolphins, orcas and octopuses) could

be classified in the rational soul instead of the sensitive soul in the hierarchy of souls. These ideas about the fluidity of the classes, could capsize the model and put strong AI even in a completely different novel ranking compared to humans.

In the end, when we take into account our understanding of a mind, Aristotle's hierarchy and by rejecting point two in Ludwig's Inconsistent Tetrad, we ought that a strong AI should be classified as a rational soul in Aristotle's classification hierarchy of souls.

# References


[1] Panpsychism. https://plato.stanford.edu/entries/panpsychism. Accessed: 2022-01-28.

[2] 'it will change everything': Deepmind's ai makes gigantic leap in solving protein structures. https://www.nature.com/articles/d41586-020-03348-4. Accessed: 2022-01-06.

[3] Seth D Baum, Ben Goertzel, and Ted G Goertzel. How long until human-level ai? results from an expert assessment. *Technological Forecasting and Social Change*, 78(1):185–195, 2011.

[4] Shell MS Weikl TR Dill KA, Ozkan SB. The protein folding problem. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2443096/, 2008.

[5] Edward Feser. Why searle is a property dualist. In *American Philosophical Association Pacific Division meeting in Pasadena, CA, March*, pages 24–28, 2004.

[6] Stevan Harnad. What's wrong and right about searle's chinese room argument? In *Essays on Searle's Chinese room argument*. Oxford University Press, 2001.

[7] Kirk Ludwig. The mind-body problem: An overview. *The Blackwell guide to philosophy of mind*, pages 1–46, 2003.

[8] Humberto Maturana, Jorge Mpodozis, and J Carlos Letelier. Brain, language and the origin of human mental functions. *Biological Research*, 28:15–15, 1995.

[9] John Searle. The chinese room. 1999.

[10] John R Searle. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424, 1980.

[11] John R Searle. Why i am not a property dualist. *Journal of consciousness studies*, 9(12):57–64, 2002.

[12] Jeffrey Skolnick, Mu Gao, Hongyi Zhou, and Suresh Singh. Alphafold 2: Why it works and its implications for understanding the relationships of protein sequence, structure, and function. *Journal of chemical information and modeling*, 61(10):4827–4831, 2021.

[13] Jason Winning and William Bechtel. Being emergence vs. pattern emergence: Complexity, control and goal-directedness in biological systems. In *The Routledge handbook of emergence*, pages 134–144. Routledge, 2019.