Inner Monologue and Machine Consciousness

Is inner monologue a sufficient criterion for machine consciousness?

Abstract

This essay has three major parts: The first parts starts by laying the foundations for a relationship between inner monologue and consciousness. I will argue why inner monologue is sufficient for consciousness (mainly by looking at the Attention Schema Theory (AST)). On this basis, in the second part criteria which inner monologue must have for machine consciousness, are elaborated, creating a catalogue applicable to architectures of machine inner monologues. The third part deals with the two questions: Whether current implementations are sufficient for machine consciousness, and if machines can have inner monologues that fulfil the criteria of consciousness in front of the background of AST. Inner monologue is used to test whether a machine is conscious in terms of AST. The result and answer suggest that inner monologue can be regarded as a sufficient criterion for machine consciousness, but that the construction and implementation of machines will not yet allow for an inner monologue sufficient for machine consciousness. This essay does not answer whether there are different forms of inner monologue that might also be sufficient, but aims to assess machine inner monologue against a catalogue with criteria for inner monologue overlapping with requirements posed by AST.

Keywords

Inner monologue, machine consciousness, consciousness, inner speech

Table of contents

1. Introduction

The explanatory gap of consciousness remains yet unanswered. There is no consensus about why or whether we are conscious, why it is something like to be and how we know that somebody else is conscious. With advances in computer science, the creation of *artificial intelligence* (AI) and maybe even *artificial consciousness* (AC) becomes increasingly probable. But how can we know if a machine is conscious? Why should it be conscious?

There are different attitudes towards the creation of *machine consciousness* (MC). The implications of an (un)intended development of a new form of conscious beings are discussed from an ethical perspective: ranging from enthusiasts to sceptics, wanting to stop the development until we have more knowledge about consciousness. It is a fact, that machine consciousness can lead to a variety of scenarios – a human-friendly machine, a suffering machine, a *super intelligence* which does not care about humanity but e.g. saving the planet and will do everything to fulfil this goal… Since it is difficult to properly define consciousness and the degree of effectiveness of a conscious machine, the implications are not foreseeable. Nevertheless, humans work on AI and *artificial general intelligence* (AGI). Many regard consciousness as the key factor for human intelligence. Thus, the development of machine consciousness is very likely if artificial general intelligence is created.

Additionally, there are components of human intelligence and human abilities which might be necessary or sufficient (or both) for consciousness. Because computer scientists try to mimic humans to facilitate and improve i.a. human-machine interaction, certain human abilities like inner monologue are mimicked in robots (Gorbenko et al. 2012, 675 ; Haikonen 2007, 148). Inner monologue is assumed to lead to more transparency, easier interaction, retraceability, understandability and consequently improved justification concerning human-machine cooperation or (moral) decisions of machines (Pipitone and Chella 2021, 13). If such a mimicked ability is sufficient for consciousness, consciousness might unintentionally be created by trying to solve another problem in the area of robotics and artificial intelligence. Others fully intend to develop machine consciousness, because they think e.g. that consciousness is necessary for collaborative machines (Graziano 2017, 7).

Inner monologue might be such an ability which could lead to the development of machine consciousness. The impact of machines with inner monologue on machine consciousness is

analysed in this essay. A lot of terms can be used equivalently (Morin 2009:1, 389): “self-talk (which includes both inner and outer self-directed speech), propositional thought, subvocal speech, covert speech, self-referent speech, internal dialogue, internal monologue, auditory imagery, subvocalizations, utterances, self-verbalizations, and self-statements”. In the following, the term inner monologue will be used to refer the phenomenon of “hearing one’s inner voice”. Inner monologue can also have a dialogic character, but is not described as inner dialogue because it rather is a form of a reflecting self (Gregory 2017, 96 ; Morin 2009:1, 394). Since an inner dialogue can also be associated with schizophrenia, the term is avoided (Morin and Uttl 2013, 3).

Inner monologue is highly subjective, and self-awareness and consciousness are often associated with subjective processes like perception, cognition and i.a. reflective thinking. Behaviourists like Watson equated thinking to inner monologue (Geraci et al. 2021, 2) and the *Language of Thought Hypothesis* (LOTH) by Fodor suggests that inner monologue is an instrument of thinking (Chella et al. 2020, 4). There also are other phenomenal aspects of consciousness like the perception of pain which does not necessarily lead to thinking and inner monologue, but this essay only deals with inner monologue as a criterion for consciousness.[1]

1.1 Recent advances in the field of machines with inner monologue

Recently, there were publications concerning the implementation of inner monologue in machines (see Chella et al. 2020 ; Geraci et al. 2021 ; Chella and Pipitone 2019 / 2021 ; Chella 2019 ; Wiggings 2018 ; Clowes et al. 2007 ; Haikonen 2007 ; Gorbenko et al. 2012). Some of them are regarded as promising architectures which can be a basis for implementing inner speech (Geraci 2021, 7). Most of the developments of self-aware robots are based on mimicking human internal states (Gorbenko et al. 2012, 675) and emulate (parts of) the human brain (Haikonen 2007, 148). Different conscious and cognitive architectures simulate or have flows of concepts or “mind objects” (Haikonen 2007, 140) which allow robots to generate their proper internal states (Gorbenko et al. 2012, 687). Architectures like e.g. the *Information Dynamics Of Thinking* (IDyOT) account for creating abstract or generic high-

[1] At this point I do not want to take either a materialists or an anti-materialist stance and discuss how thinking, inner monologue or consciousness arises.

level concepts and ideas (Wiggins 2018, 33). Others model inner monologue in cognitive architectures containing memory and phonological loops inspired by human memory (containing *short-term memory* (STM), *procedural* and *declarative long-term memory* (LTM) etc.), speech recognition and speech production which rest in a constant interactive cycle between perception, action and memory (Chella et al. 2020, 2). In this architecture, information from external perception (also called *exteroception*) are related to conscious and from internal perception (also called *proprioception*) to self-conscious thoughts (Chella 2019, 6).

As aforementioned, the advances in the field of machine inner monologue are made because inner monologue is thought to enable trustworthiness, increased human-machine cooperation and transparency in (moral) decisions (Pipitone and Chella 2021, 13). Additionally, it is claimed that for genuine self-awareness, robots need the ability to attend to their internal states (Gorbenko et al. 2012, 675).

1.2 Why inner monologue cannot be a necessary criterion for machine consciousness

Implementing inner monologue might – as previously outlined – lead to the implementation of consciousness. But is inner monologue a necessary criterion for machine consciousness?

Consciousness is defined very differently. Overall the scientific community agrees on the fact, that consciousness can be described as the phenomenal character of subjective experience (also called *qualia*). There are many other known inner phenomenal experiences like physical sensations, pure emotions, thinking without symbols and mental imagery, which are not instances of inner monologue (Chella et al. 2020, 4). Inner monologue manifests itself in different forms: from fragments of words and speech up to fully formulated and articulated sentences. (Martinez-Manrique and Vicente 2010, 3, 29). Some like Bermúdez claim that not all functions of conscious thought require linguistic thinking (Martinez-Manrique and Vicente 2010, 5). It also is a predominance of sense of meaning which also explains why inner monologue is faster than overt speech (Morin 2009:1, 393). It is not answered, if being conscious of a thought is necessarily represented by inner monologue (Martinez-Manrique and Vicente 2010, 29). Thus, inner monologue can only be a “part of the whole picture”.

Consciousness also arises and occurs in the absence of linguistic verbalisations (Morin 2009:1, 398).

Since "agentive consciousness is not supported by cognitive neuroscience and reflective introspection" (Halligan and Oakley 2021, 13), it is questionable, if the epiphenomenal character of consciousness and even inner monologue – Pompe-Alama describes inner monologue as the "feeling of what it is like to think" (Roskies 2015, 2) – necessitate one another or if they are both simply illusions (Halligan and Oakley 2021, 13).

Another argument against the claim that inner monologue is a necessary criterion for machine consciousness is, that deaf people can think and are conscious although they cannot talk to themselves and cannot hear themselves with their inner ear. Those who learn sign language start to not only use sign language for *interpersonal communication* and but also use it effectively for *intrapersonal communication* (Morin 2009:1, 395).

Further evidence against inner monologue as a necessary criterion for machine consciousness is Dr. Jill Bolte Taylor's case study. It nevertheless strongly suggests are correlation and dependence of inner monologue and machine consciousness. It is about a left hemispheric stroke after a congenital arteriovenous malformation which led to reduced self-awareness and sense of self like e.g. the perception of self-conscious emotions (Morin 2009:2, 527). Researchers have opposing views concerning this case: some highlight the role of language, others say, the damaged brain regions rather highlight the role of self-recognition and the *Theory-of-Mind* (ToM) (Morin 2009:2, 528). Nevertheless, consciousness was not completely lost although inner monologue was lost, showing that inner monologue is not necessary.

1.3 Epistemic and argumentative goal

The relation between inner monologue and conscious seems obvious. Traditionally, inner monologue received only few attention in the context of consciousness and artificial intelligence research although it is potentially regarded as a central part of consciousness by cognitive psychology and neuroscience (Haikonen 2007). This essay deals with the few investigated implications of inner monologue as a criterion for machine consciousness. If inner monologue is sufficient for consciousness, inner monologue might either be used to

determine whether machines or other beings are conscious or used to create consciousness. Hence, it is important to know what inner monologue implies in terms of consciousness:

(i) When does a being have inner monologue and what does that mean about the consciousness of the being?

(ii) Can we – and if so how – develop machine consciousness by the means of inner monologue?

The central question and epistemic goal is to investigate if inner monologue is a sufficient criterion for machine consciousness (that inner monologue is not necessary for machine consciousness has been outlined in 1.2). To answer this question positively, first inner monologue and consciousness are analysed. Then, criteria for an inner monologue, sufficient for consciousness (against the background of the Attention Schema Theory), is developed. With these criteria, it will be argued that machines cannot yet implement a sufficient inner monologue but that inner monologue might be realised in future and that this will be a sufficient criterion for machine consciousness.

2. Inner monologue and consciousness

Inner monologue is related to consciousness by different arguments. Mainly self-awareness, self-reflection, attention and thought are connecting consciousness and inner monologue. Language often was connected to consciousness, but inner monologue less. Nevertheless, there are associations can be found in the literature (Morin 2005, 118). The following sections introduce and define inner monologue, consciousness and their relation.

2.1 Inner monologue

Inner monologue is the experience of language without audible and overt talk which ranges from condensed, stripped form to expanded trains of thought (Verhaeghen and Mirabito 2021, 8). It is an interdisciplinary research field concerning disciplines like psychology, neuroscience, and pedagogy and plays an important role various cognitive processes like high-level cognition (i.a. planning, reasoning and focusing), task switching, evaluation and moral (Pipitone and Chella 2021, 1f.), memory, language acquisition (Chella 2019, 2), experience of the self (Halligan and Oakley 2021, 13), self-knowledge (Gregory 2017, 95),

self-regulation (Morin 2009:1, 389), introspection (Martinez-Manrique and Vicente 2010, 1), self-guidance, problem-solving, and awareness (Morin 2005, p.117). Inner monologue arises from a transformation of interpersonal verbal self-information to intrapersonal talk (Morin 2005, 122) at a young age and Vygotsky claims it to be crucial for cognitive and social development (Geraci et al. 2021, 2). It is a natural ability with which one can converse with oneself by either internalising other's views (Mead's mechanism) or replicating others (also called Cooley's mechanism) (Morin 2005, 122). From a social-interactionist's point of view, the self and mind arose by social communication and social complexity was first possible with inner monologue (Morin 2009:1, 400). Consequently, self-awareness and self-consciousness are required by society (Morin 2009:1, 399). By listening to and interpreting others and their external speech, humans learn about the other persons' mental states. In consequence, we acquire self-knowledge by listening to and more importantly, interpreting our internal speech (Gregory 2017, 95).

In different situations, inner monologue is used differently. Children first start to repeat language and sentences aloud before the start to think loudly and slowly internalise their thoughts (Morin 2009:1, 390). Nevertheless, adults still might use inner monologue aloud – so called *private speech* – if they are using it i.a. for concentration, structuring or regulation (e.g. "Where was I?", "Yeah, that has to go there, …, and this…") (Morin 2009:1, 389). Inner monologue can also be used when somebody is drunk or even ill. Commonly, depression, doubts, schizophrenia, anxiety or even gambler's illness have been related to negative dysfunctional inner monologue (Morin 2005, p.117 ; Morin 2009:1, 397).

An important aspect of inner monologue in terms of (self-)consciousness is awareness and self-knowledge. But a problem called the *Vulnerability Inner Speech Problem* arises (Gregory 2017, 102): "What reason, if any, is there for a subject that she believes the proposition expressed by an assertion that she makes in inner speech, given that the subject is capable of lies, deception, error and poor, ambiguous, or misleading expression?". Inner monologue seems to be an insufficient source of information. But an interesting aspect of human inner monologue is, that humans nevertheless cannot deceive themselves (Gregory 2017, 102).

2.2 Consciousness and the Attention Schema Theory

Apart from inner monologue, it is not yet known whether or why consciousness exists. In many theories, consciousness is equated with awareness, attention and *metacognition*. Metacognition is a special form of reflection used to be aware and attend to the own cognition and thoughts (Chella et al. 2020, 2). A part of consciousness is the awareness of mental content and introspection (Haikonen 2007, 148 ; Morin 2005, 117). From a psychological perspective, self-consciousness is the last developmental step of self-awareness (Chella et al. 2020, 3). In the following, awareness and consciousness will be used equivalently.

Dehaene differentiates two dimension of awareness: (i) in form of experience and (ii) of self-monitoring (Chella et al. 2020, 1). Morin, on the other hand, distinguishes three sources of awareness (Morin 2005, 116): (i) the self, (ii) the physical world and (iii) the social environment. Awareness is also suggested to be a prerequisite of the Theory-of-Mind because the *Simulation Theory* claims that people simulate the experience others make (Chella et al. 2020, 11).

Most traditional notions of "consciousness" contain the following three aspects (Hallingan and Oakley 2021, p.2): (i) the phenomenon of subjective awareness, (ii) awareness of the self and "free will" as well as control over actions, and lastly (iii) the awareness of being aware and thoughts, perception, memories, emotions. Thereby, consciousness covers similar functions as inner monologue like decision making, error-detection, optimisation, reflection, self-monitoring, learning, prioritising, control, etc. (Hallingan and Oakley 2021, p.2). Additionally, consciousness is responsible for a sense of agency (Morin 2009:2, 526).

To make the central question of the essay more decidable, a focus on the *Attention Schema Theory* (AST) by Graziano is made. The central claim of AST is that the "brain arrives at the claim that is possesses a non-physical, subjective awareness and assigns a high degree of certainty to that extraordinary claim" (Graziano 2017, 1). The difference between humans and computers lie in the phenomenal character of subjective experience (qualia) (Graziano 2017, 2). Compared to the *body schema* (where a being has a model of its physical self), AST extends the body schema and also has a model of its "own internal information-handling processes", which Graziano (2017) also calls "attention schema". In contrast to the *illusionist approach*, with which AST shares similarities, the brain also creates the illusion of qualia and

only is a mechanistic processor, but it does so not by learning (Graziano 2017, 5). Instead, consciousness is inborn because it only can make the claim to be conscious due to internal information, without which the claim would be impossible (Graziano 2017, 5-7). Consciousness is a model of the ability of deep and focused processing which is useful but inaccurate (Graziano 2017, 7).

2.3 Is inner monologue sufficient for consciousness?

By laying the foundations in the previous two sections, it is now possible to attempt an answer to the question if inner monologue is sufficient for consciousness before continuing with machine consciousness in the next chapters.

It is conspicuous that there are strong connections between self-awareness and inner monologue (Chella et al. 2020, 12). In the literature, inner monologues are a style of writing used for a long time to portray characters and their conscious experience (see e.g. Wei 2021, 335). The relationship is marked by the subjective experience of consciousness which is a central aspect of inner monologue (Morin 2005, 118) and as Sokolov claims, thought (Chella and Pipitone 2019, 1). Most thought processes also are in the area of higher-cognitive functions and self-reflection: e.g. reflection of being conscious (Morin and Uttl 2013, 7). The epistemic role of inner monologue in reasoning is both first-order reasoning and metacognition (Munroe 2021, 18), both aspects of inner monologue as well as consciousness.

AST claims that the "illusion of consciousness" is inborn which goes along with Jerry Fodor's LOTH (Morin 2009, 390). Although LOTH claims that humans think in *mentalese* – cognitive abstract mental symbols different from *natural language* (NL), there are philosophers like Wittgenstein who regard mentalese as necessary for NL (Wittgenstein 2009, 88). Furthermore, researchers like Carruthers and Jackendorff argue that conscious thought is only possible with language (Chella 2019, 2).

There also is neurological and psychological evidence that inner monologue is sufficient for consciousness. As in Dr. Jill Bolte Taylor's case (introduced in 1.2), a lack of important aspects of conscious experience and deficits in self-awareness go alongside the loss of speech, which can be e.g. caused by a destruction of Broca's area (Morin 2005, 126f. ; Morin 2009:2, 526). Another case is Helen Keller who states that by learning the langue and the meaning of

words like ‘I’ or ‘me’, she started to know that she is, and consciousness existed for her first from then on (Morin 2005, 127). Lastly, a questionnaire-study was conducted by Morin (1992), assessing inner monologue and its various functions. Between inner monologue and private self-consciousness, a significant correlation of 0.46 was found (Morin 2005, 128).

Another interesting question is raised i.a. by Churchland: “What are the empirical grounds for supposing animals are deprived of this capacity [*consciousness*]?” (Morin and Everett 1990, 352, italics added). A fundamental difference between humans and animals is language. Thus, inner monologue could be the answer since human language has a – by philosophers – controversially discussed role in the emergence of mind (Morin and Everett 1990, 352). Self-awareness remains “relatively primitive, vague and unelaborated” without inner speech (Morin 2009:1, 400).

To sum up, inner monologue shares many crucial aspects of consciousness. Thus, if inner monologue can be identified, it is sufficient to identify consciousness as well. Inner monologue can therefore be regarded a sufficient criterion for consciousness.

3. Criteria for an inner monologue, sufficient for consciousness

The previous chapter argued the sufficiency of inner monologue for consciousness. It is a first step towards answering the central question of the essay: “Is inner monologue sufficient for machine consciousness?”. By asserting that inner monologue is sufficient for consciousness, inner monologue can be now used to test for consciousness. In this chapter, a catalogue of criteria – which inner monologue must fulfil to be sufficient for consciousness and is applicable to machines – is developed.

First the criterion is outlined and afterwards, it is elaborated. Inner monologue has to fulfil the following six criteria to be sufficient for consciousness:

(i) Inner monologue has to have an attentional aspect (focusing on a specific object) in a continuous or incoherent manner, depending on the strength of stimulus and/or concentration.

Key aspect of consciousness is the awareness of the self, the experience of the self and a subjective experience which is related to different processes in the brain. It is used for decision making – a continuous train of thoughts which argues for different possibilities and arrives at a decision in a coherent manner. But thought processes can also be interrupted by e.g. loud signals or a person can be distracted (Verhaeghen and Mirabito 2021, 8). In dangerous situations a conscious thought process can also be completely interrupted, leading to instinctive reactions which are not consciously assessed before execution because a conscious evaluation would be too slow.

(ii) The content of inner monologue reflects attention/awareness influenced by first-order perceptional processes or can be steered and controlled by higher cognitive thoughts.

As AST is an extension of the body schema, it ascribes cognitive as well as linguistic access to the physical self and the model of its proper information handling processes. Furthermore, Dehaene also introduced two dimensions of awareness: awareness of experience and of self-monitoring (Chella et al. 2020, 1) which can be equated to inner monologue about perceptual experience and higher-order thoughts. Both theories stress two levels of awareness which have to be reflected by inner monologue.

(iii) The content of inner monologue has to contain talk about cognitive processes like planning, the acquisition and reflection of knowledge, control and regulation of mental processes and action, emotions, perception, action and memory.

As aforementioned, consciousness is associated with awareness of action, perception, reasoning, knowledge, … Through inner monologue these abilities and features become accessible, available and one becomes aware of them.

(iv) Inner monologue has to have a dual character: a prevalence of sense as well as (properly) articulated language.

As previously outlined, inner monologue can manifest itself in articulated language and a symbolic language similar to mentalese. It is the same for consciousness: we know that it is something like to be and additionally we can cognitively assess and think about this phenomenal experience. In AST this is equivalent to the normal processes of the body which would equate the prevalence of sense and the brain additionally creating the linguistic articulated “illusion” of consciousness “on top”.

(v) Inner monologue has to enable metacognition.

The main goal of AST is to explain that the brain only claims to have consciousness ("a non-physical, subjective awareness" (Graziano 2017, 1)) and strongly believes the claim. Since one cannot deceive oneself by inner monologue, inner monologue could be mean of transportation of the experience of consciousness. Additionally, only by inner monologue and thought can subjective awareness and knowledge about consciousness arise.

(vi) The course of inner monologue can be used to reason, justify and explain actions and decisions, thus it has to mirror conscious thought.

Graziano discusses why the brain would make up consciousness. If it does not really exist, but is only made up, the question arises, how the existence of consciousness can be justified. He claims the internal information about consciousness are useful to model the ability of deep and focussed processing. These deep processes resemble argumentative and justifying trains of thoughts in inner monologue.

4. Why machines do not have machine consciousness yet regarding the sufficient criterion of inner monologue

After creating the catalogue, recent publicised architecture of machine inner monologue can be assessed if they are sufficient for machine consciousness.

4.1 What current implementations of inner monologue in machines lack

Current implementations or cognitive architectures of inner monologue in machines strongly resemble and mimic human inner monologue (see above). They employ e.g. aspects of awareness like recognition or perception (Chella et al. 2020, 3). But they are not yet capable and able to reproduce all aspects of self-awareness, ignore attentional shifts and are very resource-intensive (Chella et al. 2020, 3). Another argument why current implementations of inner monologue in machines are not sufficient is that up to now, inner monologue was not necessary to solve problems in binary computers (Haikonen 2007, 148).

This strongly suggests that a binary computer might not be the right tool to implement inner monologue and that an implementation is not possible due to hardware. In the following section it is discussed in more detail, why the inner monologue of machines is not sufficient for machine consciousness.

(i) Inner monologue has to have an attentional aspect (focusing on a specific object) in a continuous or incoherent manner, depending on the strength of stimulus and/or concentration.

One reason why inner monologues of machines are not sufficient for machine consciousness is that they mainly implement a continuous talk which is not interrupted or disrupted and changed by attentional shifts. Attention is an important aspect of human cognition, enabling humans to cope with the great amount of incoming information. Many philosophers argue that embodiment of machines is needed for genuine consciousness (Clowes et al. 2007, 8). But this necessarily requires an attentional coping mechanism as humans have because it would otherwise be too resource-intensive and additionally not enable the required attention-shifts in the inner monologue sufficient for consciousness.

(ii) The content of inner monologue reflects attention/awareness influenced by first-order perceptional processes or can be steered and controlled by higher cognitive thoughts.

Current implementations show a basic use of language used in tightly defined boundaries (grammatical rules) (Pipitone and Chella 2021, 14). But the circuitry for inner monologue is very limited and the aspect of language grounding, which are crucial for the relation between inner monologue and consciousness (Haikonen 2007, 144), are not considered. It touches upon another problem of consciousness, the problem of genuine understanding, which is best described by the *Chinese Room Thought Experiment*[2] by Searle (Searle 1980, 418). Can a machine genuinely understand the linguistic content presented to it (Haikonen 2007, 148 ; Searle 1980, 418)?

---

[2] The Chinese Room Thought Experiment is about a person who is non-Chinese speaking and was born and raised in a very different social and cultural context. Now this person sits in a room – isolated from the world. He receives Chinese text though a window and has books with rules, telling him what to answer. He can then answer the person in Chinese according to the rules without understanding a word. Searle calls this symbol manipulation without genuine understanding, which is what a computer does (Searle 1980, 418).

(iii) The content of inner monologue has to contain talk about cognitive processes like planning, the acquisition and reflection of knowledge, control and regulation of mental processes and action, emotions, perception, action and memory.

Since the language, use of language and knowledge of the machine is fixed and hard coded in the system (Chella et al. 2020, 9 ; Pipitone and Chella 2021, 14), the scope of the inner monologue limited and thus also the possible abilities inner monologue deals about.

(iv) Inner monologue has to have a dual character: a prevalence of sense as well as (properly) articulated language.

The *hard problem of consciousness*[3] (Chalmers 1995, 3) is that there is an explanatory gap, why it is something like to be. Inner monologue displays a similar subjective experience: what it is like to think (Roskies 2015, 2). Both are not testable because of their high subjectivity. So the prevalence of sense is barely assessable.

Concerning articulated language: Since language is hard coded, the implementations and architectures are a first step, but do not come close to the articulation abilities of humans. It is nevertheless retrievable and assessable by looking at e.g. the log of a machine.

(v) Inner monologue has to enable metacognition.

Language is hard coded (see above), which leaves little room for complex processes and does not enable the machine to acquire new concepts (Chella et al. 2020, 9 ; Pipitone and Chella 2021, 14). But this also means that the cognitive scope and capability of machines is limited opposed to the unlimited capacities of humans.

(vi) The course of inner monologue can be used to reason, justify and explain actions and decisions, thus it has to mirror conscious thought.

There are implementations which already exhibit a transparent reasoning which is understandable for humans and can be used for justification purposes (Chella et al. 2020, 12). Nonetheless, since there is no metacognition (see (v)), the ability to reason and justify is limited. Furthermore, inner monologue can be equally bad assessed from

[3] The hard problem of consciousness is the subjective experience that it is something it is like to be a conscious being. The term was coined by Chalmers (1995) and opposed to the easy problems, Chalmers sees no way to explain the hard problem by computational or neural mechanisms (Chalmers 1995, 2).

an external perspective (directly or behaviourally) as consciousness (Geraci et al. 2021, 2) and machines still lack self-explanatory abilities.

4.2 Can machines have inner monologues which are sufficient for machine consciousness?

The previous paragraph showed with various arguments that machines do not yet have an inner monologue sufficient for consciousness. But will machines never have inner monologues which are sufficient for machine consciousness? Many problems will be solvable with continuous development and research. But some problems which are especially hard, are listed now:

Embodiment as a probably important factor to both inner monologue and consciousness has been introduced before. Inner monologue arises though and probably arose because of socialisation and interaction whereas consciousness also is linked to embodiment. *Hesslow's simulation hypothesis* argues that agents have inner worlds in which they can simulate external-world interactions (Clowes et al. 2007, 10). A self-model which could be portrayed and formed by inner monologue. (Clowes et al. 2007, 2). It is questionable if inner monologue and machine consciousness can be created/are created in a new, foreign form to human inner monologue and consciousness, otherwise proper embodiment must be realised which copes with the high amount of information without much resources.

Another problem might be the binary mode of operation of machines. Because it did not necessitate and support the implementation of inner monologue, another form of computation could be needed to first be able to make advances in the area of inner monologue.

Lastly, language grounding poses a big challenge for inner monologue development.

5. Conclusion, implications and outlook

The essay discussed the relation between inner monologue and machine consciousness, arriving at the conclusion that inner monologue is sufficient for consciousness and also machine consciousness, although not for current architectures and approaches of machine inner monologue. Of course it has to be considered that AST which was used as a foundation to assess machine consciousness, receives criticism and is also highly debated. Thus, the

result of the analysis also has to be regarded critically. Further research is needed to close the explanatory gap of inner monologue: Why is it like something to think? If this can be answered, a machine inner monologue is potentially thinkable in future. Whether it is worth striving for conscious machines was not discussed. But because it is a hotly debated topic, inner monologue could be an important aspect. Should conscious machines have inner monologues and if so, why? Why should inner monologue be an important factor in future machine consciousness research?

Graziano (2017, 8) claims that consciousness will be necessary for machines to develop empathy, thus encouraging the development of machine consciousness. But since it is difficult to detect and identify consciousness in non-human entities, inner monologue could be a good tool, pointing at consciousness and facilitating testing. There are various arguments used to argue why inner monologue would be useful to machines, especially in the area of social interactions, trust and transparency (Chella et al. 2020, 2). It would lead to many of the listed abilities like awareness, better adaptation, control, or regulation and maybe even the creation of a Theory-of-Mind (Chella et al. 2020, 3,5). Furthermore, new machines should act autonomously and self-governed (Geraci et al. 2021, 1) which requires high cognitive abilities. Because they become collaborative agents of the world (Laird et al. 2017, 15f.), people have to trust them and they have to act according to human expectations (Graziano 2017, 7 ; Pipitone and Chella 2021, 13), especially since *ethopoeia*[4] will take place (Geraci et al. 2021, 5). Thus, inner monologue would definitely enrich machines. But in terms of consciousness it could also make the behaviour of machines more transparent, less unpredictable and retractable, especially if ethically difficult decisions are concerned (which would face the problem of *AI explainability*[5] and AI as a *black box*[6]) (Geraci et al. 2021, 6 ; Haikonen 2007, 148 ; Pipitone and Chella 2021, 13).

Inner monologue could also be the missing part in the puzzle to create machine consciousness, because language is regarded as a crucial aspect of the human brain and mind

---

4 The attribution of attittudes, intention or motives to non-human entities (Geraci et al. 2021, 6).

5 The problem with artificial intelligence in machine learning contexts is that it is impossible for humans to understand and retrace how the intelligence arrived at the result it did.

6 “Black box” is a term used to describe that no one understands and knows what happens inside one concrete calculation in machine learning.

(Haikonen 2007, 148 ; Roskies 2015, 4) and a mimicry of human-like minds is strived for by some researches (Clowes et al. 2007, 12).

In humans, inner monologue and consciousness go hand in hand and language and socialisation play an important role connecting them. As there are unsolved questions, a definite answer is not yet possible. But it seems likeable that conscious machines will also exhibit some form of inner monologue.

References

(1) Chalmers, D. J. 1995. Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies* 2(3):200-19 (Chalmers 1995)

(2) Chella, A. 2019. A Cognitive Architecture for Inner Speech. *Cognitive Systems Research*. 10.1016/j.cogsys.2019.09.010 (Chella 2019)

(3) Chella, A. / Pipitone, A. 2019. The inner speech of the IDyOT, Comment on "Creativity, information, and consciousness: The information dynamics of thinking" by Geraint A. Wiggins. https://www.sciencedirect.com/science/article/pii/S1571064519300247, last access: 20th February 2022, 7:51pm. (Chella and Pipitone 2019)

(4) Chella, A. / Pipitone, A. / Morin, A. / Racy, F. 2020. Developing Self-Awareness in Robots via Inner Speech. *frontiers in Robotics and AI* 7:16. doi.org/10.3389/frobt.2020.00016 (Chella et al. 2020)

(5) Clowes, R. / Torrance, S. / Chrisley, R. 2007. Machine consciousness: Embodiement and imagination. *Journal of Consciousness Studies*, 14:7, 7-14. (Clowes et al. 2007)

(6) Geraci, A. / D'Amico, A. / Pipitone, A. / Seidita, V. / Chella, A. 2021. Automation Inner Speech as an Anthropomorphic Feature Affecting Human Trust - Current Issues and Future Directions. *frontiers in Robotics and AI* 8:620026. doi.org/10.3389/frobt.2021.620026 (Geraci et al. 2021)

(7) Gorbenko, A. / Popov, V. / Sheka, A. 2012. Robot Self-Awareness: Exploration of Internal States. *Applied Mathematical Sciences*, 6:14, 675-688. (Gorbenko et al., 2012)

(8) Graziano, M. S. A. 2017. The Attention Schema Theory: A Foundation for Engineering Artificial Consciousness. *frontiers in Robotics and AI* 4:60. doi.org/10.3389/frobt.2017.00060 (Graziano 2017)

(9) Gregory, D. J. 2017. *Inner speech: A philosophical analysis*. Thesis submitted for the degree of doctor of philosophy at the Australian National University. https://philpapers.org/rec/GREISA-8, last access: 4th March 2022, 1:43pm. (Gregory 2017)

(10) Halligan, P. W. / Oakley, D. A. 2021. Giving Up on Consciousness as the Ghost in the Machine. *frontiers in Psychology* 12:571460. doi.org/10.3389/fpsyg.2021.571460 (Halligan and Oakley 2021)

(11) Haikonen, P. O. A. 2007. Towards Streams of Consciousness; Implementing Inner Speech. *Integrative Approaches to Machine Consciousness, 5th-6th April 2006*, 144-43. researchgate.net/publication/242321244_On_Architectures_for_Synthetic_Phenomenology, last access: 10th March 2022, 7:27pm. (Haikonen 2007)

(12) Laird, J. E. / Lebiere, C. / Rosenbloom, P. S. 2017. A Standard Model for the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. *AI Magazine* 38(4). doi.org/10.1609/aimag.v38i4.2744 (Laird et al. 2017)

(13) Martinez-Manrique, F. / Vicente, A. 2010. What The...! The Role of Inner Speech in Conscious Thought. *Journal of Consciousness Studies* 17(9-10),141-167. (Martinez-Manrique and Vicente 2010)

(14) Morin, A. 2005. Possible Links Between Self-Awareness and Inner Speech. Theoretical background, underlying mechanisms, and empirical evidence. *Journal of Consciousness Studies* 12(4-5), 115-134. (Morin 2005)

(15) Morin, A. 2009. Inner Speech and Consciousness. *Encyclopedia of Consciousness* 1, 389-402. (Morin 2009:1)

(16) Morin, A. 2009. Self-awareness deficits following loss of inner speech: Dr. Jill Bolte Taylor's case study. *Consciousness and Cognition* 18, 524-529. 10.1016/j.concog.2008.09.00 (Morin 2009:2)

(17) Morin, A. / Everett, J. 1990. Inner speech as a mediator of self-awareness, self-consciousness, and self-knowledge: An hypothesis. *New Ideas in Psychology* 8:3, 337-256. (Morin and Everett 1990)

(18) Morin, A. / Uttl, B. 2013. Inner speech: A window into consciousness. dx.doi.org/10.12744/tnpt.14.04.2013.01 (Morin and Uttl 2013)

(19) Munroe, W. 2021. Why are you talking to yourself? The epistemic role of inner speech in reasoning. *Noûs*. 2021; 1-26. doi.org/10.1111/nous.12385 (Munroe 2021)

(20) Pipitone, A. / Chella, A. 2021. What robots want? Hearing the inner voice of a robot. *iScience* 24, 102371. doi.org/10.1016/j.isci.2021.102371 (Pipitone and Chella 2021)

(21) Roskies, A. 2015. Thought, Language, and Inner Speech. A Reply to Ulrike Pompe-Alama. *Open MIND*: 33(R). doi.org/10.15502/9783958570801 (Roskies 2015)

(22) Searle, J. R. 1980. Minds, brains, and programs. *The behavioural and brain sciences* (1980) 3, 417-457. (Searle 1980)

(23) Verhaeghen, P. / Mirabito, G. 2021. When you are talking to yourself, is anybody listening? The relationship between inner speech, self-awareness, wellbeing, and multiple aspects of self-regulation. *International Journal of Personality Psychology* 7, 8-24. doi.org/10.21827/ijpp.7.37354 (Verhaeghen and Mirabito 2021)

(24) Wei, Z. 2021. Functions of Inner Monologue and Free Association in Andrei Bely's "Petersburg". *Advances in Social Science, Education and Humanities Research* 588, 331-335. (Wei 2021)

(25) Wiggins, G. A. 2018. Creativity, information, and consciousness: The information dynamics of thinking. *Physics of Life Reviews* 34-35 (2020), 1-39. doi.org/10.1016/j.plrev.2018.05.001 (Wiggins 2018)

(26) Wittgenstein, Ludwig. 2009. Major Works, Selected Philosophical Writings. New York, London: *Harperperennial*. (Wittgenstein 2009)

Fonts

Bodoni 72, Tiimes New Roman

Index & Abbreviations

Appendix

Statement of Authorship

I hereby certify under oath that the paper I am submitting is entirely my own original work except where otherwise indicated. I have not used any auxiliary means other than those listed in the bibliography or identified in the text and any use of the works of any other author, in any form, is properly acknowledged at their pint of use with indication of the source.

Anaïs Siebers (firstname, surname)

15.03.2022, Grenoble (date and place)


(signature)