What is a mark of the mental? What are some candidate marks of the mental and how do they relate?

Final essay for the seminar
"Introduction to the Philosophy of Cognitive Science"

Winter term 2022/2023
Eberhard Karls Universität Tübingen

by

Luca Anna Kosina
Matrikel-Nr.: 6084778

Introduction

Even before the first occurrence of computers, machine learning or artificial intelligence, humans have been wondering about mental states and intentions of systems outside of their own species. We are curious about the difference in animal minds, we try to differentiate levels of consciousness and try to judge affective states of other systems. Through the recognition of the possibility to create artificial computer systems with an ability to learn autonomously we now also attach judgments to the will of such artificial systems - for instance, questioning whether their goals interfere or align with human intentions. More fundamentally this comes down to answering the old question of mindedness in other systems: how do we define and judge the mindedness of artificial systems? We want to call factors for the assessment of this, marks of the mental and assess commonly named ones as well as their relationship. While John Searle argued against the mindedness of artificial systems because of their lack of intentionality as a mark of the mental, we want to criticize this in regard to its relation with consciousness. I will argue for a differentiation between categories of marks of the mental and assess another common mark of the mental, intelligence or rather intelligent behavior and its relationship to teleology.

Marks of the mental and the Chinese room experiment

John Searle has challenged the claim that machines can truly understand language or have mental states through his Chinese room thought experiment: if a person is given a set of rules of turning Chinese symbols into English language and vice versa, when passing questions about a Chinese text to the person and receiving English replies we would assume understanding of the language. Knowledge about how the answers were produced by the "translator" however, shifts the view from the person comprehending the semantics and turns it into pure syntax manipulation (Searle, 1980). We can transfer this to artificial systems: they might give us an output that implies understanding of semantics, but a closer look into the "black box" of the system reveals its content to be formal symbol manipulation. Searle comes to the conclusion that artificial systems are not minded if we assume that the act of language production and the process of manipulating symbols are not as signs of mindedness, since "purely formal principles you put into the computer will not be sufficient for understanding, since a human will be able to follow the formal principles without understanding anything" (Searle, 1980, p. 187).

What is it, however, that differs the understanding of a human? Searle argues that "[...] the formal symbol manipulations by themselves don't have any intentionality. They are meaningless—they aren't even symbol manipulations, since the 'symbols' don't symbolize anything." (Searle, 1980, p. 199) This intentionality, that is the directedness of the content of mental states, such as thoughts, and also what we communicate, would be a factor differentiating a conscious human from an artificial system like a computer or even strong artificial intelligence. Searle argues that this manipulation alone cannot generate intentionality or consciousness - the argument is not only about the presence of understanding or intelligence in machines, but also about the nature of consciousness and subjective experience. Searle claims that not all intentional states are conscious and vice versa (Searle, 1983). But without the possibility of consciousness would intentional states even be a mark of the mental? Without conscious experience of a language, would humans also not just manipulate syntax?

The conscious experience itself has been argued to be a mark of the mental often before in the philosophy of mind. René Descarte famously argued that the mind is a non-physical substance and only the subjective experience of itself, of being aware of one's thoughts and feelings, proves the existence of the self (Descartes, 2008). It is this knowledge about the possibility of my own conscious experience of, for example, beliefs which are intentional, that lets us make inferences about the mental states of other systems in Searles argumentation. Without knowing what understanding of semantics means as a conscious being, how would we be able to differentiate the intentionality of symbols for one system and the lack of the same for another? Searle additionally assumes that in "[...] cognitive sciences one presupposes the reality and knowability of the mental in the same way that in physical sciences one has to presuppose the reality and knowability of physical objects." (Searle, 1980)

However, I want to object to this reasoning since one of the challenges in determining whether artificial systems can be minded is the problem of being unable to judge other systems' intentionality and consciousness. We can never directly experience another system's consciousness or phenomenology, and therefore must rely on indirect evidence and inference to determine whether it exists. Searle assumes there to be a fundamental difference in understanding of humans compared to other systems and intentionality to be the necessary mark of the mental. However, we could not differ between the intentionality of, for example, a human and a thermostat if it were not for the assumption of consciousness in one system. If intentionality and consciousness were necessary marks of the mental, but we can not be sure

about the ontological status of the consciousness of the mental states of other humanes, we could not assume that other humans are minded. This again relativates why we would even assume intentionality to be necessary for mindedness - if we do not even know if there is a difference between someone understanding Chinese and someone following rules for symbol manipulation.

What this should make clear is that assessing mindedness through intentionality and consciousness as marks of the mental relies on unaccessible phenomenal experiences. I want to differentiate this from assessing marks of the mental from an intersubjective perspective. This corresponds to David Chalmer´s discrimination of the hard and easy problem of consciousness who claims that we cannot explain phenomenal consciousness in terms of function (Chalmers, 2010). Just as explaining cognitive abilities and functions differs from explaining phenomenal consciousness, we can see a difference in categories of marks of the mental. The explanatory gap shown by Chalmers therefore affects our judgment of mindedness and while we can still ask the question if it is possible that performance of these cognitive functions can take place without any subjective experience, we should not only view mental states which are conscious to a subject as marks of the mental - mental states can be present without having a system that corresponds to what we think makes the human mind special. Instead we can also look at marks of the mental for which might judge the mindedness of a system from the outside or intersubjectively.

One such alternative approach to the problem is to use intelligence or rather intelligent behavior as an indicator of the latter. If a system behaves in ways that suggest it possesses consciousness, then we may infer that it does. For example, if an artificial system is able to respond appropriately to unexpected or novel situations, solve problems and communicate solutions, it may suggest that it possesses some higher level mental states. Especially the ability to produce language and communicate with other system should be suggested as mark of the mental - from the perspective that we can only know the product of the language understanding and production process of a system - just as Wittgenstein claimed that subjective mental states are irrelevant for the judgment of the status of an entity if it demonstrates the ability of verbal communication. This is also the foundation of the Turing test in which intelligence of a dialogue partner is judged by a subject having conversations with the system.

Teleology of intelligent behavior as mark of the mental

One can object that intelligent behavior alone is not sufficient to determine mindedness: it is possible for a system to behave intelligently without actually possessing mental states comparable to a human or even animal. For example, a chess-playing program may be able to beat a human at chess, but intuitively we would not ascribe mental states to it. This is different, though, for other, stronger, artificial intelligence systems and we should therefore find a mark explaining the gradual difference of mindedness of systems.

We can ask what is is that differs a system like a chess-program from the actions a human takes during the game of chess and what seems obvious at first glance, is that the human can transfer its ability to think logically from chess to other problems, they can also behave more flexible, whereas the program is constrained by its set of rules. In contrast to a chess-playing program the human playing chess has agency for his decision to play chess and awareness of his goals in the game, whereas a program might simply respond to inputs and follow a set of rules or algorithms. We can introduce the concept of teleology to describe this difference in systems: Teleology refers to the idea that things have purposes or goals, and that these purposes or goals can guide behavior. In the context of the mental, it can be seen as a mark of agency, or the ability to act with intention and purpose. A system with teleology is able to change in its environment, not simply reacting to inputs, but actively seeking out opportunities to achieve its goals and using its knowledge and resources to accomplish them, which, for example, a chess-program can not do. A higher-order artificial intelligence or natural language processing model can use new inputs to learn and evolve - it might therefore be attributed to awareness of its environment - in contrast to, for example, a simple search engine.

Important for why teleology could be a relevant mark of the mental is its emergence from the system´s ability to set itself into relation to the world around it - it realized itself as subject and at the same time object in the world. So it does interfere with the concept of self-consciousness, but in contrast to intentionality it is not a mark of a mind only from the perspective of the mind itself, but on the other side: it is characteristic for the interaction of the system with the world.

Additionally, it gives us the chance to view mindedness not as polar, but to assess it as “gradually” on a scale: there seem to be different stages of intelligent behavior and different levels of autonomy when it comes to teleology, which according to Searle is, for example, not the case for intentionality.

In summary, while Searle has presented an argument for intentionality and consciousness to be fundamental for mindedness, these turn out to be unhelpful. We instead argue for the differentiation of subjective marks of the mental and such that we can apply intersubjectively to other systems. For the latter, commonly proposed is intelligence or the observation of intelligent behavior. However, this alone does not seem to satisfy all situations in which we assess the mindedness of another system. One potential additional mark of the mental therefore proposed is teleology of the intelligent behavior.

References

Chalmers, David: The Character of Consciousness, Oxford University Press, 2010.

Descartes, René: Meditations on first philosophy, Oxford University Press, 2008.

Searle, John: ‘Minds, Brains and Programs’, *Behavioral and Brain Sciences,* 1980.

Searle, John: Intentionality: An Essay in the Philosophy of Mind, Cambridge University Press, 1983.