Why Robots Cannot Feel

In recent years, the efficiency of artificial intelligence has been impressively demonstrated. In scenarios whose states and changes are fully definable – such as in the games Chess and Go – AI systems are now far superior to humans. However, neural networks capable of learning, which – following the example of evolution – permanently optimize themselves by selecting the most successful variants, achieve considerable success in areas of the real world too.

So it is understandable that the hopes of AI now go much further: Is it possible to create a system that equals or even surpasses human performance not only in specific areas, but also *in total*? Can an information processing system be constructed that has *consciousness*?

In any case, there seems to be no *absolute* obstacle for the realization of this vision. Obviously, also the brain itself is an information-processing system. And this applies also to all sub-structures of the brain, including those that are necessary for our feelings – they all are nothing other than biological modules that receive information in the form of electrical impulses, process it and pass it on to other structures.

So if one assumes that it is precisely this information processing in our brains that creates mind and consciousness, then it seems obvious that it is only *technical difficulties* what separates us from creating a robot with consciousness – albeit on such an enormous scale, that even within AI there are gradually increasing doubts that the construction of such a robot will be possible in the foreseeable future.

Here, we will ask ourselves whether it is really only technical difficulties what prevents the creation of a conscious machine or at least postpone it to a distant future, or whether there are also obstacles *in principle* – and by that I mean obstacles that *in no way* can be eliminated.

Let us assume we have succeeded in constructing a robot that has an artificial neural network whose structure corresponds to that of a human child. This neural network is supplied with information from the outside world and from the body of the robot via artificial sensory organs in the same way as in a human. In the function that simulates the connections between the neurons, we have implemented all the changes that occur in natural neural networks, i.e. the amplification through activity and the reduction through non-activity, and also the modulation of these connections through chemical systems. This seems to ensure that the robot is capable of *learning* in the same way as a human: it will have a *memory*, it will form *representations*, it will be able to *think*, etc.[1]

Let's call our robot ***Joe***.

How will Joe evolve? Will he have feelings? Will he develop consciousness?

Given the above conditions, it actually seems natural that the answer has to be: ***Yes, he will***.

Yet this answer is wrong. Rather, the following is true:

Even if Joe were the best possible simulation of a human, he would feel nothing and would have no consciousness.

Why is that? The proof is surprisingly short and simple.

First we define simulation:

---

1 The prerequisites of the thought experiment are intentionally so extremely idealized, because the only question here is whether our project will fail even if *all* technical problems have been solved. So the robot *should be* a perfect simulation. (For that, the list of his skills is still rather incomplete.)

"Simulation" is the reconstruction of the dynamics of a really existing system in another system constructed for this purpose.[2]

In contrast to the "replica" of a system, the dynamics of the simulation is not caused by the same driving force as the dynamics of the original system.

For an illustration, let us look at simulations of our solar system. In earlier times, mechanical simulations were very popular – often beautiful constructions in which balls made of wood or brass imitated the movements of the planets around the sun. Today we will rather find computer simulations in which suitable algorithms generate a video of these movements.

In any case, *it is not gravity* what drives the simulation – as is the case in the real system. And it is immediately evident that it can never become gravity, no matter how much the accuracy of the simulation is increased. Obviously, gravitation as driving force of the dynamics would only be preserved in a *replica* of the solar system. (In this replica, the representations of the celestial bodies would have to appear with the masses of the originals!)

The dynamics of a system is based on the *causal relationships* through which the objects of the system are linked to one another. For the construction of a simulation it is therefore necessary to determine the *causal level* of the system, i.e. the level on which the processes take place that cause the dynamics of the system.

In the solar system, this is trivial, since there is only one single "level": the objects are the celestial bodies, their movements are caused by gravity.

In the human neural network, on the other hand, we find three levels: the physical, the neural and the mental level. In my paper *The Substantiation of Free Will*, the *mental level* has been determined as the *causal level*. I will briefly repeat the reasoning:

The physical level: Here an enormous number of processes run simultaneously, many of which influence one another. Therefore, there is absolutely no method for predicting the future development of the network. The assertion: "What happens in the network follows from initial conditions and physical laws" is wrong. The same applies to the neural level.

The mental level: Neural patterns that represent or mean something can be produced by the network without an external cause. They must therefore be understood as *attractors* of the network.[3]

An attractor determines the dynamics of a system if the state of the system lies in the catchment area of the attractor.

The state of the neural network of a human is *always* in the catchment area of an attractor: from *any* state, the network will immediately adjust to a pattern that *means* something.

So it can be stated:

In the human neural network, the mental level is the causal level. Mental processes determine the dynamics of the network.

Now we have to ask:

What is the driving force behind the dynamics of the mental area? What drives us to think and act the way we do?

---

2 *Dynamics* means the development of the *state* of a system; *state* is the totality of the attribute-values of the objects of the system. (E.g. their positions and momentums.)

3 *Attractor* is a system state or a sequence of system states – so to speak a "pattern", towards which the system *necessarily* evolves and which it then maintains for a certain period of time.

The answer is:

Sensation.[4] **Sensation is the driving force of the dynamics of the mind.**

Since the mental area is the *causal* area of the neural network, it follows:

Sensation is the driving force of the dynamics of the human neural network.

Previously, we have established that exactly *that* what drives the dynamics of a really existing system, is *not* transferred to a simulation of this system. If we now apply this fact to the simulation of a human neural network, then we get:

When a simulation of a human neural network is carried out, sensation is not transmitted.

This means:

In the simulation, there is no sensation but only information.

And here, too, applies what we previously found in the simulation of the solar system regarding gravity: No matter how far the accuracy of the simulation is increased – what drives the dynamics of the simulation will never become sensation.

In other words:

The simulation – the robot – does not feel anything. Our robot Joe is not a sentient being but a zombie.

If sensation is absent, then there is no consciousness either: Even the most abstract intellectual activity is carried by an interest and guided by a motive, and both interest and motive are descendants of sensations from which they cannot be separated. So it would be quite absurd to ascribe consciousness to a robot without sensation.

This is the answer to the question why robots will *never* have sensations and consciousness.

Heinz Heinzmann

August 2021

4 *Sensation* must be understood here in the broadest possible sense: It stands for everything that goes *beyond* information in a mental state, i.e. for that which can not be *defined* but only be *experienced*. (Two examples: the frequency of the color red can be defined, but the sensation *red* can not; the strength of a pressure can be defined, but the sensation *pain* can not.)