Why Free Will Exists and Why Robots are Not Sentient

Contents

1. The Substantiation of Free Will

Abstract

1. First, the difference between reality and description is determined. Based on this, it can be shown that the physical causality – in the following referred to as "causality from below" – is *incomplete*.

2. This is a necessary condition for assuming causality in more complex layers of reality governed by nonphysical laws. This type of causality – in the following referred to as "causality from above" – is explained by an example and then generally justified.

3. The explanation applies also to the human neural network. From this follows that the mental layer is the *causal layer* of the network.

4. In contrast to the laws of physics, mental laws are changeable. Since the mental processes are the causal processes, also these changes must be attributed to the mental activity.

5. Therefore, to a voluntary decision the following applies:

a) It is not a *physical* but a *mental* process.

b) The decision-making process can change the laws that applied before it started. However, if only by this process itself is decided what will happen, the decision cannot be determined beforehand.

So it is free.

1.1. The Difference Between Reality and Description

In our universe, the following seems to apply:

***Everything that exists* consists of elementary objects that interact with each other. How these objects behave *is completely regulated by physical laws*. Thus, the entire future development follows from so-called "initial conditions" – the totality of the attributes of all objects at any point in time – and physical laws.**

In this picture that is so convincingly presented to us by science, there seems to be no room for anything other than physics. No matter how complex the aggregates are into which the elementary physical objects are assembled, no matter what fantastic creations evolution produces – *ultimately* everything remains physics. There is just no room for anything else.

This fact can be specified as follows:

In this so-called *reductionist* view of reality just presented, causality always remains "below", i.e. in the *elementary layer* of reality. All other, more complex layers have lost their independence. Descriptions that refer to these layers – such as neural or psychological descriptions of human actions – are just simplified, approximately valid summaries of processes that are actually of physical nature.

The consequences of these hypotheses are rather strange, if not to say bizarre. If we assume, for example, that we made an assertion B *because* it is logically correct, then that would be a self-deception: It would mean postulating a causality at the level of mental processes, so to speak a causality from "above" – which, however, is inadmissible after what has just been said; B would then be "causally overdetermined". If this "causality from above" could actually claim an independent existence – *in addition* to the "causality from below" – then it would have to be possible to decide *against* the physical causality.

There would be only one possibility that B could actually correspond to logic: that evolution had adapted the physical processes in our brain to the requirements of reality to such an extent that we behave and think logically to a sufficient degree for our survival. But I emphasize again: the conviction that we made the assertion B *because* it is logical would be a delusion, a ruse of evolution to reinforce our adapted behavior through a pleasant feeling. And, incidentally, we would never be able to determine whether something like "logic" exists at all, since *understanding* something would also be a mental process that does not exist *as such*. Insights would not be insights, thoughts would not be thoughts, mind would have disappeared, *we ourselves* would have evaporated in the fog of self-delusions ...

So it is a completely absurd picture that follows from the reductionist view, and I believe that it is only so widespread because no reductionist has ever fully considered the consequences of his or her convictions. (If there still were one, however, he or she would have long since fallen silent and would therefore be untraceable.)

I want to briefly touch on the two most popular attempts to "defuse" the problem.

The first objection is, that – because of quantum mechanical uncertainty – in nature itself an "objective indeterminacy" exists, so that it cannot be said that "the future follows from initial conditions and laws". However, it can be said that "the future depends *exclusively* on initial conditions and laws" – save that these laws are no longer deterministic. The following conclusions then remain valid.

However, the most common objection to reductionism is, that in most cases a complete reduction has not been achieved and will probably never be possible. I consider this objection inadequate: whether there *is* a reduction cannot be decided by whether *we* are able to carry it out – the picture of

reality sketched above, which is the basis of the incredible success of natural science, is not questioned by the restrictions which *our* means and abilities are subject to, and this applies also to the conclusions drawn from this picture.

Therefore, in order to avoid these strange inferences, it is necessary to put the picture itself into question. So we ask: *Is the hypothesis A true?*

A: *Everything which happens follows from physical laws and initial conditions.*

Let us start with a thought experiment:

We consider the following scenario: a large number of any material objects in empty space that are moving randomly relative to each other, but in such a way that they remain gravitationally bound to one another.

Let us assume that we were able to grasp the initial conditions – the totality of the attributes of all objects of the system – with absolute precision and transfer them to a description. So we ignore that we cannot measure with infinite accuracy, or that we are not even able to write or store the value of a single attribute with infinite accuracy. We also assume that our law of gravitation is correct and that we are able to perform all the necessary calculations.

Now we compare the situation in the *really existing system* with the situation in the *description system*.

Under the above conditions, in the *existing system* exactly what we expect will undoubtedly happen: every object will behave precisely as gravitation dictates. Thus, here, hypothesis A seems to be confirmed.

And in the *description system*? Well, here, at first *nothing at all* happens. Although we have inserted the infinitely precise values of all attributes into our equations, so that they actually represent the objects and their development in time *perfectly,* still the equations do not behave like the objects themselves: While – starting from the point in time that we have chosen to measure their attributes – the *actually existing objects* move on *by themselves* and, in this way, carry out the gravitationally determined dynamics of the system, the *equations* obviously do *not* do that – they simply remain unchanged as we have noted them.

This is actually completely obvious. Nevertheless, I was a little more explicit than necessary because here we have come across an extremely important issue, which, however, so far almost completely escaped both philosophical reflection and scientific research – presumably precisely *because* of its ostensible obviousness. It reads as follows:

<u>Proposition:</u>
There is a fundamental difference between a really existing system and its representation: the really existing system is <u>active</u>, but the representation is <u>not active</u>.

Let us return to our thought experiment. We have stated: In the *existing system*, every object will behave exactly as gravitation dictates. Does this actually confirm hypothesis A?

The answer is: *No, it does not!* Actually, we have *added* something to the really existing system that is not contained in A: *activity.*

The fact that reality is *active* means: at any point at any time exactly what has to happen happens *by itself*. It means that reality doesn't have to *calculate* anything, that it doesn't need a law or an algorithm, because it simply processes all individual cases at the same time.

Obviously, however, *activity* is precisely that which cannot be transferred from the reality to its representation. It can be said that the *type* of activity of the system, its *specific structure*, must be contained in our equations of the gravitational field, but the *activity itself* is missing.

Let us note: Because of its *activity*, reality advances *by itself* from the present to the future. But the description system refuses to do us this favor. In order to obtain information about the future of the system in our description, we therefore need a *mathematical procedure* that substitutes the missing activity.

Do we have such a procedure? First of all, it is clear that for a "large number" of objects that move randomly, our equations cannot be solved. In fact, we have only one way to obtain knowledge about the further development of the system: Since we know the gravitational field, we can calculate for each object where it *would have moved* after a certain time interval *in this field* – and here, the subjunctive is necessary because of course it does *not* move in *this* field: indeed not only the object we are looking at is moving but also all other objects, and this means that also the field itself is constantly changing. But in order to be able to calculate anything at all, for small time intervals we have to assume the field as *static*. We then do the same kind of calculation for all bodies. Then we repeat this procedure for the next time interval etc.

The crucial point is that from start to finish we depend on *approximations*, and that we also do not know to what extent our calculations deviate from reality. At the latest after the next branching point – that is a point in the development of a system at which an arbitrarily small difference in the initial conditions can lead to completely different states of the whole system – our prediction becomes pure luck.

With this we have shown that hypothesis A is false. Since there is no procedure which enables us to conclude the future from the present, A cannot be maintained.

Proposition:

There are systems whose future development does <u>not</u> follow from physical laws and initial conditions.

But isn't reality itself constantly showing us that the future follows from the present? Not at all. What we see is just that the future "follows" the present. It is only this suggestive picture of reality conveyed by physics that leads us to believe that everything "follows from" initial conditions and laws. However, the expression "follows from" is a logical conjunction that can only relate to a description. To apply it to reality means to replace the "follows" that we observe with the "follows from" that we postulate; But we have to *justify* this act of substitution, and so we are forced to replace our "follows from" by a series of logical steps. Thus we inevitably end up with a mathematical procedure, and finally again with the fact that no such procedure exists – even if we imagine we were freed from all restrictions of measuring and calculating.

So the future does not always follow from the present. What does this result mean?

The most important consequence is that a *logical free space* is created: If initial conditions and physical laws were sufficient to derive the future, then there would be no room in the set of conditions for the derivation of the future; But since they are *not* sufficient, there is now room for further elements in this set.

Proposition:

Causality from below is incomplete. There is room for causality from above.

1.2. Non-Physical Causality

Our next step will be to clarify what kind of "further conditions" could exist on which the future development of systems depends – in addition to initial conditions and physical laws. Is it any other kind of data? Or other kinds of laws? To determine this, we change the scene.

We consider a simple glass vessel. When we hit it, it vibrates and makes a sound. What does this tone depend on? What determines its height and character? The answer is: *the shape of the vessel*. It gives rise to a mathematical law that enables us to predict the vibration pattern of the glass. So here we don't have to go into the physical objects – the glass molecules – nor the physical interaction – the electromagnetism – in order to predict the sound. The only physical information needed is the speed of the sound propagation in the glass.

The law that now allows us to predict the future of the system is therefore *not a physical law*. It belongs to another kind of laws which I shall call **Laws of Form** or **Laws of Structure**.

Let us compare our two scenarios, that of the gravitating bodies and that of the vibrating vessel:

In the gravitation scenario, the initial conditions are given as **local parameters**, as attributes of the individual bodies. Their values are inserted into the **physical law** – the law of gravity. Although everything that happens fully conforms to this law, it is still impossible to predict the further development. The future of the system **does not follow** from its present.

In the glass scenario, it is not the attributes of the glass molecules that are inserted into the law, but the dimensions of the glass, i.e. **global parameters**. The law is not a physical law, but a **Law of Structure**. The further development can be derived from the global parameters and the law. The future of the system **does follow** from its present.

The sound that we hear is largely independent of the way we produce it. However, this does not apply to the first moment: initially, there is a transient process that depends on how we strike the vessel. Only after this process it does always vibrate in the same state. This state to which the glass ultimately adapts – the vibrational pattern into which it develops and which it then maintains – is called **attractor**.

Above, we asked ourselves what types of data and laws could there be in addition to physical initial conditions and laws. The simple example of the vibrating vessel gave us an answer:

1. new data in the form of *global parameters*.
2. new laws in the form of *Laws of Structure* that are based on the global parameters.

Since these new data and laws can be used to predict the future of the system, they are in fact elements of the "set of conditions for deriving the future" mentioned above.

However, most important for our considerations is undoubtedly the following:

The local parameters – such as the positions and velocities of the glass molecules – initially depend on where, with what and how hard we hit the vessel. So at first they can be quite different. Regardless of this difference, the state of the vessel always evolves towards the same vibrational pattern – the attractor.

In the case of a glass vessel, there is only one possible vibration pattern that always develops, regardless of how the vessel is struck. The future movements of the components of the vessel – the glass molecules – are therefore determined by this pattern.

Causality works from the whole to the individual, from the vessel to its components, and not the other way round.

Proposition:

A form of "causality from above" occurs when in a system *attractors* exist, i.e. states which the system will *inevitably* evolve into, if it is "close enough" to the attractor state.

(A necessary condition that it is actually "causality from above" is that the physical causality in the respective system – the "causality from below" – is *incomplete,* just as we have demonstrated in the gravitation scenario. However, since the glass vessel was only intended to demonstrate what our argument is about, we do not need to worry about whether this condition is met here.)

Now we have made all necessary preparations to move on to our final and decisive scenario:

1.3. The Human Neural Network

Subject of our investigation is the following question:

What kind of causality does the neural network obey?

In the network, there are three levels of increasing complexity:

1. the physical level
2. the neural level
3. the mental level

In relation to this classification, our question is:

Of which kind of processes does it depend what happens in the net? Of physical, neural or mental processes? Which level is the <u>causal</u> level? – Or, to put it another way: Which level is <u>dominant</u>?

First to the physical level. Let us assume we had complete knowledge of the values of the attributes of all physical objects in the network and could thus set up the system of equations that represents the state of the network and its further development. (Of course this idea is completely absurd, but in the form of a thought experiment it is permissible – *in principle*, this system of equations must exist.)

But now we are again confronted with the problem that already prevented the calculation of the development of the system in the gravitation scenario: An enormous number of processes are running at the same time, and each of them is directly networked with several others. In order to be able to calculate any process, we have to assume at least for a small time interval that its immediate environment is constant – i.e. we have to isolate it for a short time. Then we can do the same for all other processes, and after that we repeat the whole procedure for the next time interval etc.

As with the gravitation scenario, we are therefore dependent on approximations that can deviate considerably from reality already after a short time. It is not possible to predict how the network will develop. The claim "What happens in the network follows from initial conditions and physical laws" is wrong.

And here, too, the following applies again: Reality does what we are not able to do: due to its *activity*, it executes the enormous number of processes at the same time, so that we get the impression that everything "follows from" initial conditions and physical laws.

6

Proposition:

In the neural network, the physical causality is incomplete. There is room for causality from above.

Let us now consider the *neural level*. It consists of many billions of neurons. Each neuron is directly connected to hundreds or even thousands of other neurons, and *all* neurons are linked to one another via a few intermediate steps.

The neural activity is regulated by a law that follows from the neural input-output mechanism.[1] This law can be understood as the *law of interaction* of the neurons. (It also serves as basis for computer simulations.)

Also at the neural level, it initially seems completely natural to us that what will happen in the network follows from the initial conditions of the neurons and their law of interaction. And again we have to recognize that we succumbed to the same deception, in that we have not differentiated between reality and description or confused them:

Since the neural interaction law is a summary of physical circumstances, the argument with which we have just refuted the claim that everything follows from initial conditions and physical laws remains valid. Thus for the neural level the following applies: The high degree of networking of the neurons – the permanent feedback that results from it – precludes the existence of a mathematical method for calculating the further development.

Proposition:

Also the description by neural initial conditions and the neural interaction law leaves room for causality from above.

This brings us at last to the most complex level, the *level of the mind*. We make the following assumptions:

1. Every kind of mental activity (thoughts, chains of associations, sequences of images, etc.) is a sequence of neural activation-patterns.

2. Sequences of neural activation-patterns can be representations of facts.[2]

Let us look at the neural patterns. How do they become representations?

Let us imagine a neural network in which there are no representations yet. An object perceived for the first time will cause a certain pattern in this network, starting from the primary visual cortex. The neural connections that are active are strengthened because of this very activity. The same is the case with each repetition. This gradually creates a stable connection between the object and a specific neural pattern (or rather an ensemble of specific neural patterns).

In addition, the following applies: Although the neural patterns are initially caused by external stimuli, after a sufficient number of repetitions they are also produced by the neural network independently of these stimuli. This means:

Neural patterns that are connected to objects in the manner just described are* attractors *of the network. (See also the associated notes here and here.)

---

1 The expression "input-output mechanism" means the following: The dendrites of each neuron are stimulated or inhibited by other neurons via synapses. The electrical excitation caused in this way is passed on to the cell body and added up there. When a certain limit is exceeded, it is released to the axon and distributed to its branches, so that ultimately it influences other neurons via synaptic connections.

2 Here, "facts" must be understood in the widest-possible sense.

Previously we have stated:

Under the condition that the causality from below is incomplete, from the existence of attractors follows that the respective system – provided it is "close enough" to the attractor state[3] or in this state itself – is governed by causality from above.

However, according to our first premise, a mental process consists not only of neural patterns, but also of the transitions between these patterns. But to this transitions the same applies as to the patterns themselves: First, they are determined by the sequence in which the causative objects appear. If this sequence is repeated, the corresponding neural activity is reinforced, and this has the consequence that the patterns occur again in the same sequence even if they are generated by the network itself. In the same way, also the spatial relationships of the objects are transferred to the patterns.

This means: In the processes that are generated by the network itself, the neural patterns that are in a stable connection with specific objects appear in the same spatial and temporal contexts as the objects themselves. Therefore, *the patterns can be understood as representations of the objects, and the processes as representations of the facts in which the objects appear.*

So, in human neural networks it is not the physical or neural conditions and laws that determine what happens in the network, but *the structure of the network* – the fact which attractors there are and how their sequence is regulated – on which the processes depend that run in the network.

Causality acts from the whole to the individual, from the network on its components, and not the other way round.

We have thus achieved our first goal:

Proposition:

The neural network is regulated by *causality from above*. The mental level is the dominant level. In it lie the *causes* for the processes running in the network.

So the statements we made so far were *actually* conclusions and not just physical processes! Or – to follow up on the formulations used in the criticism of reductionism: Insights are insights, thoughts are thoughts, mind is set in its rights, *we ourselves* are indeed we ourselves …

So far, so good, but that doesn't take us to where we actually want to be. Just because we have moved causality up doesn't mean we are free. We have only replaced physical or neural causality with mental causality. We have thus achieved that our mind is not ruled by physical or neural laws, but *by its own law: the Law of Structure, which the sequence of neural patterns obeys that represent something*.

But don't we ultimately remain trapped in the scheme of initial conditions and laws from which we wanted to escape? Fortunately, that's not the case. To show this, we need to look at the difference between physical and mental laws.

1.4. The Difference Between Physical and Mental Laws

Human neural networks differ greatly from one another, even if they have not yet been structured by external stimuli. From this follows immediately that the patterns that represent something are also different in all people, even if the represented facts are identical.

---

3   Without the concept of phase space, this "close enough" cannot really be defined. In any case, the neural network is always "close enough" to an attractor state.

As stated above, initially the order of the patterns is determined by the order in which the objects or circumstances that cause the patterns occur. But as soon as the network itself is able to produce these patterns, the transition rules of the patterns – what we have called the *mental law* – increasingly depend on their use in internal processes. This dependence on external and internal conditions means that the transition rules differ from person to person.

So we have already determined the first difference:

*While physical laws are **generally valid**, mental laws are **individually valid** – they only apply to one singular person.*

Connections between neurons are strengthened when they are active,[4] and weakened when they are inactive. This means that every mental activity alters the structure of the network. But if the structure can change, then obviously also the rules that determine the sequences of the neural patterns can change.

So this is the second difference: *Physical laws are **immutable**, mental laws are **modifiable***.

Proposition:

Physical laws are universal and immutable. Mental laws are individual and modifiable.

1.5. The Substantiation of Freedom

The most obvious implication of the strengthening of active neural connections is that what we always think, feel and do is self-reinforcing. Basically, however, it goes without saying that also the opposite can occur:

We have shown that causality is to be found at the mental level. *Will* and *intention* must be understood as elements of mental causality. Now let us imagine concretely we were faced with an important decision. When we enter the decision-making process, we are initially guided onto certain, well-known paths by the regularities that are valid up to that point – i.e. by our own mental law.

But at any time we are able to leave these paths, for example by simply considering the opposite of what we have assumed up to then, or by taking a path we never tried before; We are able to do so precisely for the reason that the causes for what happens in the network – and thus also for the modifications of the network structure – lie at the mental level.

In other words:

The law that determines the sequence of neural patterns in our network that represent something, i.e. our own mental law, can be altered *by ourselves*: we ourselves can change the laws of our thinking and acting through our thinking and acting, and we can do it *deliberately*.

This means at the same time:

Although mental processes are governed by their own rules, it is not possible to derive a volitional decision from them: the decision cannot be contained in these rules because they can be changed by the mental process that precedes the decision. While this process is taking place, the laws that it obeys can change – or, more precisely, *it itself* can change the laws that applied before it started.

---

4   This finding goes back to Donald Hebb, who stated in *The Organization of Behavior* in 1949: "When an axon of cell A is near enough to excite B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased."

<u>**Proposition:**</u>

Volitional decisions are causes of actions. Since only by the decision-making process itself is decided what will happen, the decision is not determined beforehand.

So the decision is free.

To the question of why a (sane) person has decided so and not otherwise, there is then only one permissible answer:

Because he/she wanted it that way.

<u>Note:</u>

Of course this does not mean that volitional decisions cannot be analyzed with respect to their neural, chemical, physical, genetic, social etc. causes. It means, however, that these analyses necessarily remain incomplete and never lead to a secure result, because mental phenomena cannot be reduced to other layers of reality. The will remains the final authority.

Postscript

In reviewing the text, it seemed to me that I followed my goal of presenting the topic as briefly and simply as possible perhaps a little too radically. Therefore I will try to explain the most important points of my argument one more time:

Let us assume we have to describe a system that consists of a large number of physical processes that are linked to one another. Then the equations of the processes are also networked with one another. For an exact description, we would therefore need the values of all parameters of any process at every moment in order to insert them into the equations of all other processes – in other words: it is (except in very simple cases) *impossible* – for reasons of principle, and not just because of the limitations of measurement and calculation – *to make accurate predictions* about the system that consists of all these processes *by using physical means*.

And with that we would have actually reached the end of our possibilities – *unless* the processes could be understood as elements of a "structure of a higher order", in which further laws apply. These "higher order laws", however, are then *no physical laws*, and with that we have left the field of physics.

If these new laws make it possible to predict the development of the overall system, then the following applies:

1. The development of the overall system ***does <u>not</u> follow from physical laws***.
2. The development of the overall system ***does follow from higher-order laws***.

Of course, everything continues to happen ***in accordance*** with the laws of physics – but these laws now take place within a ***higher-level structure***. (Think of the vibrating <u>glass vessel</u>.)

So causality is no longer *below,* which means: in the elementary, physical realm. It has migrated *upwards*, into a realm of higher order, in which ***new, non-physical laws*** apply.

Exactly these conditions can be found in the neural network, and in fact several times:

In a neuron, numerous physical processes take place at the same time. Although the physical approach allows us to understand what is going on in the neuron, still the coupling of the processes prevents any exact calculation of the further development. However, due to the shape and structure

of the neuron, these processes are embedded in a system of higher-order, so that they obey a "structural law" – the one that we previously called "neural input-output law".

Now, however, it again applies that this law does not allow us to make any precise predictions about the future development of the many neurons that are coupled to one another. But the neurons themselves are again elements of a higher-order system: the neural network with its imprinted patterns (attractors). So the neurons are also subject to a new law: a structural law of again higher order: the law of the sequence of neural patterns, and that means: *the law of the mind*. Thus mind is the causal layer; It determines the processes that take place in the network – including those that change this law itself.

Finally, I shall repeat the difference between description and reality:

In order to get from the present to the future in the *description* of a system, we need some kind of procedures. These can be mathematical procedures, algorithms or equations, but also methods to combine facts in such a way that conclusions can be drawn. In some cases we are able to do this so well that we can state: *B follows from A*.

In the *reality*, none of this is necessary. If what has to happen happens in every place at every time, then the future will arise *by itself*, and then all complex objects and structures, including their laws, will develop *by themselves*.

But from the fact that in the reality the execution of elementary processes is sufficient for the creation of the future, it cannot be concluded that the future *follows from* elementary processes, because that would presuppose that that, what in the reality happens *by itself*, can be expressed by *a series of logical steps*, and that is impossible.

Note:

In this justification of free will, it is not necessary that a "bifurcation" exists in the development of the world. The key point here is that *the future is not contained in the present* – that is, it does not *follow* from the present but merely *arises* from it, and that the reasons for what will then actually happen are of a mental nature.

Note:

In order to recognize objects, artificial neural networks must be trained on large data sets. In numerous repetitions the connection strengths of their neurons are varied until a sufficiently high recognition rate is achieved.

In contrast, we started from the following hypothesis: A perceived object, which causes a neural activation pattern, is represented *by this pattern itself*. Therefore, here the relationship between object and representation is not established by varying the connection strengths of the neurons, rather it exists already from the beginning and is only stabilized and specified by *strengthening* the active connections, whereby the neural pattern becomes an *attractor*.

This hypothesis is confirmed most clearly by the so-called "imprinting". (As e.g. in the case of the gray geese of Konrad Lorenz). There are neither "large data sets" nor "numerous repetitions" – the process occurs almost instantaneously.

Furthermore, thereafter *immediate recognition* occurs, despite the inevitable variability of the sensory impression to be recognized. Thanks to the attractor concept, this – otherwise hardly explainable – performance becomes self-evident: as long as the sensory input is within the catchment area of the attractor, it obviously applies: *perceiving = recognizing*, since the newly activated attractor already represents the object, so that further calculations are unnecessary.

2. Why Robots are Not Sentient

Preliminary Note

The content of this section follows in part from the statements made in the previous one. However, because of the current importance of the topic, I consider it suitable to carry out the proof in full. So I will present the necessary facts and arguments here again, albeit shortened.

I have decided on a two-stage implementation: for the first, short version of the proof, the expansion of the scientific view presented in the Section on Free Will is sufficient: there we have freed the mental level of reality from the grip of physical causality, by showing that the *activity of reality* cannot be imitated by logical or mathematical procedures, so that the claim that everything **follows from** physical initial conditions and laws cannot be maintained.

Under this assumption, it is possible to understand mental states as *self-dependent, dominant objects*, which is made concrete by conceiving them as attractors of the dynamics of the neural network. The sequences of these states – the mental processes – can thus be determined as the *causal layer* which this dynamics depends on.

However, in order to secure the proof against all possible refutations, it is necessary to analyze the scenario in more detail and to reconstruct it conceptually. It is then not enough to shift causality "upwards" – into the mental realm – but rather the complete concept of reality is required, according to which reality is *more* than a describable sequence of states that (in principle) could be reproduced with any degree of precision.

2.1. First Version (Short Form) of the Proof

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 (and fears) 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 it is currently uncertain whether 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 or delays the creation of a conscious machine, or whether there are also obstacles *of 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.[5]

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**:

"Simulation" is the reconstruction of the dynamics of a really existing system in another system constructed for this purpose.[6]

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!)

Therefore applies:

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

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.

---

5   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.)

6   *Dynamics* means the development of the *state* of a system; *state* is the totality of the attribute-values of all objects of the system at any given point in time.

In the human neural network, on the other hand, we find three levels: the physical, the neural and the mental level. In the Section on 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.[7]

However, the following applies:

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 claimed:

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?

The answer is:

Sensation.[8] **Sensation is the driving force of the dynamics of the mind. Information without sensation is indifferent and therefore passive.**

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.

---

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

8 *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 cannot be defined but can only be *felt* and *experienced*. (Two examples: the frequency of the color red can be defined, but the sensation *red* cannot; the strength of a pressure can be defined, but the sensation *pain* cannot.)

In other words:

The simulation – the robot – does not feel anything. It cannot love or hate, want or not want. 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 absurd to ascribe consciousness to a robot without sensation.

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

2.2. Ontological Expansion and Validation of the Proof

Although the short form of the proof we have just performed is complete, it has a weakness: Since it is not entirely clear why the proof works, it might appear that it would not include an AI system, whose structure is sufficiently similar to the structure of a human (or animal) neural network, if that system was realized through *hardware* and not just through software on a conventional computer.

In a mechanically or electrically powered simulation of the solar system, we *know* that the movements of the bodies are *not* caused by gravity, and it is completely self-evident to us that the mechanical or electrical drive can never turn into gravity.

Why do we know this? Because we have a clear concept of the connection between *mass* and *gravity*, or, to put it even more clearly: of their *inseparability.*

But in the case of mental activity in a human neural network, comparable self-evident knowledge is missing. However, we already have the necessary elements:

In deriving free will, we showed that causality is not to be found in the physical processes, but in the mental activity. With this we have left the physical realm. But this "leaving the physical realm" cannot be limited to causality, but rather affects the entire description of the system. So the objects we are now analyzing are no longer molecules or neurons, but *mental states*, and the processes are not physical, chemical or neuronal, but *mental* processes.

It is therefore also clear that exactly what causes the dynamics of the mental activity cannot be of a physical kind, but must be of a mental nature.

In the short form of the proof we have determined *sensation* as that which drives the dynamics of mental activity.

Thus from this perspective, *sensation* has the same status in the realm of mind as *mass* in the solar system, so that we know with certainty that that, what drives a simulation of mind, can never become sensation.

However, the problem is that this way of looking at things is so unusual or even strange that it lacks the self-evidence we can presuppose in the case of gravity.

If we assumed, for example, that the processes in the solar system and the processes in a simulation of the solar system were almost completely identical, then we would still be convinced that the simulation is *not driven by gravity.* In general terms: for us, the approximate identity of the dynamics of the original and of the simulation is by no means synonymous with the identity of the two systems themselves.

On the other hand, exactly the same situation is completely unclear in the case of mind and its simulation: *At present **no one knows** whether the approximate identity of the dynamics of mental*

activity and its simulation also means that in the simulation sensation and consciousness are present.

However, according to our analysis, this is certainly *not* the case. So, from our point of view, it is simply the lack of a concept of mind that causes this ambiguity, and therefore it seems appropriate to now go into the ontological foundations of this question in more detail.

We start with the difference between reality and description that we introduced in the Section on Free Will:

Really existing objects are <u>active</u>, but objects in a description are <u>not active</u>. Thus, the existence of real objects must include something that objects in a description <u>lack</u>.

This element of the existence of real objects we call **substance**. Therefore, <u>**substance is that, from which the <u>activity</u> of existing objects emanates**</u>.

The element of the existence of real objects that we can perceive and describe is *the way in which they are active*, i.e. their behavior and their effects.

This element of their existence we call **accidents**. Natural science deals *exclusively* with accidents.

Therefore the following **proposition** applies:

Really existing objects consist of substance and accidents, whereas objects in a description consist exclusively of accidents.

Since an object cannot *cease* to be active in its characteristic way, ***substance and accidents form an inseparable unity***. (The earth exists only **with** gravity.)

For us, every existing object consists of these two elements: of ***substance*** – that is that part of existence whose "being there" we recognize as necessary, but which can neither be imagined nor described as what it actually "is", and of ***accidents*** – this is the part of existence that can be described and defined.

In the physical realm of reality – or let us say: in the realm of matter – these conditions are familiar to us. We know that *mass* causes gravity and that *electric charge* causes electromagnetic interaction. So we know that *there must be **something*** that is the cause of the dynamics, and we name it, but we don't know what it "is".

Now we have to determine what is to be understood as substance and accidents in the realm of the mind. Since we are no longer in the physical realm, we cannot simply use the systematization that applies there. Rather, the objects of the mental reality must first be defined, and then it must be determined what their substance and accident are.

In the Section on Free Will we stated:

Every mental state is a neural activation pattern. These patterns are attractors of the dynamics of the neural network. Every mental process is a sequence of such patterns.

These statements concern the question of how the objects and processes of the mental realm can be understood in relation to their *material presuppositions*.

But now it is our task to grasp them for what they are as *mental phenomena*. The answer is as follows:

Every mental state is a combination of two disparate elements: <u>*information*</u> *and* <u>*sensation*</u>.

Its ***information*** content is what it *represents* or *means*.

For the determination of *sensation*, I repeat what was said in the short form of the proof:

Sensation stands for everything in a mental state that goes **beyond information**, i.e. for that *which cannot be defined but can only be **felt** and **experienced***. (Two examples: the frequency of the color "red" can be defined, but the sensation *red* cannot; the intensity of a pressure can be defined, but the sensation *pain* cannot.)

(I will refer to mental states as **qualia**. The term *quale* therefore stands for the entire mental state and not just for the feeling part.)

With the above determinations, it is also clear what the substance and the accident of the mental state are:

Information is obviously that which is accessible to our thinking – that which can be *defined* and *processed*.

Therefore *information processing* is the accident of the quale.

Sensation, on the other hand, is that which *cannot be defined*, that is, that which eludes our thinking and our descriptions.

Therefore *sensation* is the substance of the quale.

And from this follows, as we already stated in the short form of the proof:

Sensation is what drives the dynamics of the mind.

Now we are sufficiently prepared to carry out our proof formally and thereby secure it.

First we need the following **definition**:

What an object is due to the inseparable unity of its substance and accidents, we call its <u>essence</u>. The activity that results from this unity we call <u>essential</u>.

(Thus the **essential activity** of the Earth is to exert gravity.)

The purpose of this definition becomes immediately clear when we now turn to *simulations*.

For example, consider a mechanical simulation of the solar system in which the model bodies are moved through mechanical devices – chains, gears, shafts, etc. – in this way mimicking the movements of the celestial bodies.

The *essential activity* of the model bodies would obviously be to exert gravity. But it is *not the mass of the model bodies* that drives the dynamics of the simulation – that is, what causes the desired movements – but *the mechanics we have constructed,* which must then be *activated*, electrically or mechanically (e.g. by turning a crank).

To express this point, we will refer to this type of activity as **supplied activity,** in contrast to the just defined **essential activity**, which happens **by itself**.

With this, the *definition of simulation* given in the first version of the proof changes in the following way:

The dynamics of a simulation is <u>not caused by the essential activity</u> that arises from the <u>inseparable unity of substance and accidents</u> of the objects of the simulation, but <u>by supplied activity</u>.

The accidents from which the dynamics of the simulation is formed are *without substance:* the substance of the objects of the simulation *is not the substance that belongs to these accidents* and with which it forms an inseparable unity, but only their *material basis* from which these accidents can be separated at any time. (As is immediately apparent in the mechanical simulation of the solar system.)

The final building block of our proof is the following **proposition**:

As long as accidents of higher complexity can be described as functions of accidents of lower complexity, the associated substance remains the same. If this functional connection is broken, the substance changes. *For us* it then appears as a new, second substance.

Before we turn to proving this proposition, we must clarify to what extent accidents in more complex layers of reality can be described as functions of accidents in simpler layers.

For example, the processes in neurons can be described as functions of the physical and chemical properties of these neurons. (Which does not mean, however, that they can be *calculated*.) The same applies in principle to all evolutionary transitions: from the physical to the chemical level, then to the biochemical, cellular, neural level, and even up to the realm of simple neural networks that do not produce mind: the processes taking place in such networks can be described as functions of their architecture and external conditions.

Only at the very last of these transitions – the transition to neural networks that produce mind – does the chain of reducibility end:

As we established when substantiating free will, then the following applies:

Initially the order of the neural activation patterns is determined by the order in which the objects or circumstances occur that cause the patterns. But as soon as the network itself is able to produce these patterns, the transition rules of the patterns – what we have called *mental law* – increasingly depend on their use in internal processes.

This means that the dynamics of the neural network – i.e. *the mind* – increasingly decouples itself from the causal chains of the environment and instead develops its own internal laws. And from this follows that the information content – i.e. the *accident* of mental states – can no longer be represented as function of the accidents of the underlying layers of reality.

Now to the proof of the above proposition.

(The totality of physical accidents we will call *first accident*, their associated substance *first substance*, the totality of mental accidents *second accident* and their associated substance *second substance*.[9])

We have just established that the accidents of all evolutionary levels can be traced back to the accidents of the levels below, with the exception of the accidents of the highest, i.e. the mental level.

It applies:

Substance and accident always form an *inseparable unity.*

The *first accident* is *inseparably* linked to the *first substance*.

---

9 However, that does not mean that there are two substances – rather, the second substance is thought of as emerging from the first substance, and the question we ask ourselves is therefore: Why does the first substance in the case of qualia *for us* transform into the second substance sensation?

If complex accidents can be reduced, step by step, to simpler accidents, then this means that they can ultimately be reduced to the first and simplest accident.

For us, however, *reducibility* is tantamount to *ontological identity*: if B is reducible to A, then B **is** actually A. So if a complex accident is reducible to the first accident, then it **is** actually the first accident, and then it is *inseparably bound* to the first substance.

Thus as long as the accidents are reducible, the associated substance remains the same – it is then still *first substance*.

But if the chain of reducibility to the first accident is interrupted by the appearance of a new, *irreducible* accident, then this new accident differs from the first accident and from all other accidents that can be derived from it.

However, due to the ***inseparability*** of first substance and first accident, the following applies:

*If the substance of an object is the **first substance**, then the associated accident must be the **first accident**.*

And from this follows:

If an accident appears that is different from the first accident, then the associated substance must also be different from the first substance.

Here is a sketch for illustration:
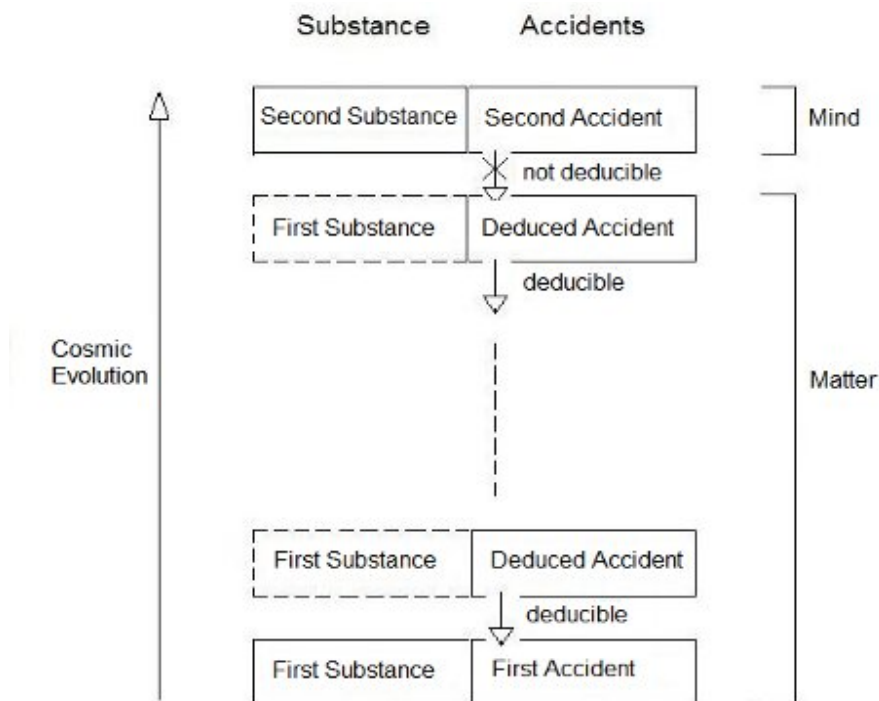
The crucial point in our argument is that the transformation of the essence of being can only occur if the dynamics of the system arises from the *inseparable unity* of substance and accidents. *Only then* does the transformation of the associated substance follow from the fact that the accidents are no longer reducible to the first accident.

If, on the other hand, the dynamics of the system is based on *supplied activity*, then the accidents are *without substance*, and the substance that belongs to the objects of the system *does not form an inseparable unity with these accidents*.

And this means:

There is no reason for the transformation of this substance. It remains *first substance.*

In other words:

The *essence* of the simulation remains *physical*. The simulation remains an information processing system without sensation.

The metamorphosis of matter into mind does not take place.

The just mentioned condition that the dynamics of the system must arise from the *inseparable unity* of substance and accidents, does not only apply to the last, i.e. the mental level – it must be satisfied on *every* level that develops during the evolutionary rise from matter to mind. If on any of these levels the dynamics of the system is not caused by the *essential activity* of the objects but by *supplied activity*, then the unity of substance and accidents is torn and the transformation of the essence of being can not occur.

So what does this mean for our proof that robots cannot have consciousness?

For AI systems that are implemented using software on conventional computers, the proof is valid without exception: the use of software is always associated with supplied activity.

But what about a *replica* of a biological neural network that reproduces the neural (analog-digital) input-output law using suitable hardware and whose structure corresponds to the structure of the entire network, so that it could be assumed that the sequence of states of the *constructed system* would almost be identical with the sequence of states of the *biological system?*

Could the transformation into sensation take place here?

The answer is clearly **no**. The condition for the transformation is not met: the dynamics of the replica is *not caused by **essential activity** but by **supplied activity***.

*The problem is that from the usual scientific view of reality **this fact cannot be understood at all***.

In this view, reality is *equated* with a (describable) sequence of states, and it must therefore be expected that the increasing convergence of two sequences of states ultimately leads to the identity of the systems themselves.

However, in the expanded materialist view that we have presented here, the concept of existence is augmented by an element that takes us beyond the realm of the describable.

This means that all our descriptions and ideas about the processes in nature are necessarily *incomplete*. So to speak "behind the scenes" of the part of the stage that is accessible to us, something happens, which is either completely hidden from us or can only be recognized and understood through inference from the part of reality that is accessible to us: the accidents.

Here, reality is *more* than a sequence of states.

In the context of our considerations, this implies:

From the approximate identity of the state sequences of the natural system and the artificial system cannot be concluded that also their essence is approximately identical.

Concretely: The substance of the two systems can be quite different despite the extensive identity of their states: [10]

In the **biological system**, the substance is **inseparably bound to the accidents of the system** and is therefore **transformed into the mental substance sensation.**

The **constructed system**, however, is driven by **supplied activity,** and therefore here the substance stands in a merely constructed and *by no means inseparable* connection with the accidents of the system, so that it **remains physical substance and is not transformed into sensation**.

The result of our conclusions is as follows:

<u>Proposition</u>:

It is not possible to construct a robot that experiences sensations and has consciousness. Neither in a <u>simulation</u> nor in a <u>replica</u> of a system that produces mind can the transformation of matter into mind take place.

There is no ghost in the machine.

Thus only *artificial intelligence* can be constructed and not *artificial mind*.

Does this mean that it is impossible to create artificial mind at all?

No. Our argument only excludes the possibility that mind can be *constructed*. However, the definition of the term *replica* can be expanded to include artificial evolution, i.e. an evolution that is designed and controlled by us.

In this case – just as in natural evolution – the condition could be met that the respective system activity is always *essential*. If we do not intervene at any point in this artificial evolutionary process through constructions or by supplying activity, but limit ourselves to controlling and accelerating the development, then at the end of this evolution there *could* be a system that produces mind.

However, no one can know whether such an artificial evolution is possible, or whether the path that nature has chosen is the only viable one.

In any case, it is clear that the creation of artificial mind remains a very distant, perhaps never achievable future, if it is not impossible at all.

<u>Note:</u>

Everything that can be defined is attainable through information processing, *everything that can not be defined* remains unattainable for it: no matter what function is applied to information – the result will always be just information and nothing else; the information "red" will never turn into the sensation *red*, the information "pressure" will never turn into the sensation *pain*.

Therefore, "information" and "sensation" (as we used it <u>above</u>) form the *only* pair of concepts that makes it possible to draw a clear and definite line between artificial intelligence and human mind and to provide evidence for it.

From this follows that the concept "consciousness", which is often at the center of the discussion, is only suitable for drawing this boundary, if the mental phenomena attributed to it (in its respective definition) are analyzed and classified according to their affiliation with *information* or *sensation:*

---

10 Of course, here "identity of the states" can be replaced by the (much weaker) condition "identity of the output" of the two systems, which represents the criterion in the Turing test. The Turing test is therefore not suitable for determining whether AI systems are sentient.

the part of consciousness that belongs to information processing (e.g. any kind of self-representation) can be reproduced – no matter what technical difficulties stand in the way of its simulation, while the part that belongs to sensation remains inaccessible to AI.

It would therefore be an unnecessary and misleading complication to base the difference between AI and mind on the concept "consciousness".

Note:

As stated at the end of the Preliminary Note, shifting causality "upwards" is not in all cases sufficient for our proof. The reason for this is as follows:

Let us assume that a neural network could be constructed that is capable of forming and connecting attractors – just as we assume for human neural networks,[11] and that this attractor network would be the *causal level* of the system. Nevertheless, the system would remain *insentient:* the condition that its dynamics is based on **essential activity** – that is, that it arises from the *inseparable unity of its substance and accidents* – would not be fulfilled.

Note:

To the hypothesis that objects are represented by attractors, the following should be added:

The pattern that forms in the primary visual cortex as the consequence of a perceived object, is not *as such* transferred directly into the neural network. Rather, it is broken down into several components – in this sense *parametrized* – which, at the end of the whole visual data processing, are assembled to the overall neural pattern that we understand as attractor.

This parametrization is an important aspect of the attractor hypothesis: The attractor is defined by a subset of the phase space. The *attractor state* of the system corresponds to a trajectory that does not leave this subset for a certain period of time. However, already a (small) subset of all according parameter values – which do not even have to be very accurate – is sufficient for restoring the attractor state, which means: a *fraction* of the complete original sensory input is sufficient for recognition. This makes recognizing objects extremely easy and, at the same time, increases the ability to generalize objects and facts.

Note:

Finally, a comment on the scenario of the gravitating bodies at the beginning of the section on free will:

Even a Laplacian demon with infinite resources of space, time and information could not carry out the calculation: In order to *accurately* determine the future of the system, the demon must perform the calculation for infinitely small consecutive time intervals. If the interval boundaries are as close as the *real* numbers, the calculation will not be finished even after an infinitely long time, but if they are *less* close (like the rational numbers, for example), it will happen that an instability is missed that occurs *between* two time points of his calculation.

In fact, even with this argument, we have still not grasped the full extent of the problem: We have assumed that – because we possess complete knowledge of the initial conditions – we know the gravitational field. However, this assumption is wrong for the following reason:

---

11 The currently so popular artificial neural networks (e.g. GPTs) are not suitable for generating attractors.

Let us denote the point in time at which we have precise knowledge of the initial conditions – and at which our calculation should begin – by t_0. If we want to calculate for any of the bodies, let's say for body A, where it will move in the first time interval, then we must know all effects from the other bodies which A is exposed to at time t_0.

For example, let's look at body B: we know the position where it is at time t_0. However, the effect originating from B that A is exposed to at time t_0 does **not** originate from **this** position, but from a position where B was **before** – exactly as long before as it took gravity to move **from there** and reach body A at time t_0. Therefore, in order to determine the effect of B on A at time t_0, we have to put B on its path *into the past*, and exactly the same applies to all other bodies: they all have to be put into the past – the further, the further they are away from A.

This means: Before we can even **begin** to determine the path of A, we first have to determine the paths of all other bodies. But for that it is necessary to also know the effect that A has on the other bodies at time t_0, and therefore we also have to shift A itself on its path into the past, i.e. on the path that is *not known to us*, since we just wanted to calculate it!

The same applies to *every* body: in order to shift it into the past, we must know the paths of all other bodies. However, since we do not know *a single one* of these paths, it is impossible to determine the exact positions where the bodies were before, and therefore it is also impossible to determine the effects which they are exposed to at time t_0.

In other words, we – and by "we" I mean all of us **and** Laplace's demon – are not only unable to **perform** an **accurate** calculation of the future, we are even unable to **begin** with it.

The scenario is not computable. *Reality* is not computable. *We ourselves* are not computable.

[So the formal version of our ontological argument about free will is as follows:

The behavior of all elementary objects is determined exclusively through physical laws. But if you try to derive the future (or, if objective chance should be factored in: *any* version of the future) in a physical way, you fail because it would require an uncountable number of logical procedures.

In some cases, however, the uncountable set of logical procedures can be replaced by a finite set of statements about a higher, *non-physical* level of reality. The facts which these statements refer to can then be understood as causes (or *reasons*) for the future state.]

3. What follows for AI?

3.1. Introduction

The usual question is:

"Why is there something indefinable in the mind, like 'color' or 'pain', and <u>nowhere else</u>?"

We asked ourselves the question instead:

"Why does the indefinable, which exists <u>everywhere in reality</u>, change its character when it appears in the mind?"

So the question is not about the reason for the *existence* of this indefinable – which would be superfluous because its existence is self-evident [12] – but about the reason for its *change*.

In the first version, the question cannot be answered. In this (false) form, it leads to strange hypotheses, such as *qualia eliminativism* or *panpsychism*.

As we have shown, however, in the second version the question can be answered, and this answer also contains proof that ***sensation*** – the *mental manifestation* of this "indefinable" – ***does not exist in systems that did not arise through evolution, but are constructed by us***.

So AI systems are not sentient. This sets a fundamental limit on the expectations, hopes and fears of AI engineers. However, it is not yet clear what this proof means for the potential performance of AI systems. In this third part of the paper, we will therefore address the question of what limitations artificial intelligence is fundamentally subject to due to its lack of sentience.

3.2. What is *"sensation"*?

First, a brief comment on why I determined the term "sensation" differently from its usual meaning:

Every mental state contains something that is ***not definable***, which goes ***beyond information***. However, since there is no term which all possible elements of mental states can be assigned to – and because I wanted to avoid a long list – I have instead chosen the term that comes closest to this missing term: ***sensation***.

Therefore, on the one hand here the term "sensation" is restricted compared to its common use (because it is supposed to contain *no information*, i.e. no *definable* part), but on the other hand it is also significantly expanded.

Two examples were used to illustrate this: *color* and *pain*. Color, because the indefinability of the color sensation is a known fact, and pain, because it is perfectly understandable that the event "hammer blow on finger" triggers a mental state that contains not only the information "hammer head is in contact with finger" but *something more*: the sensation *pain*, which can be so strong that it is impossible to deny its occurrence.

"Sensation" understood in this way can be divided into three areas:

A) The first area is the area of *perception*:

Sensation encompasses the entire "inner theater": the virtual space, the stage on which we act, which is always present to us <u>as a whole</u> – as an "image" – and on which we see, hear, feel, smell and taste.

---

12 See Part 2, <u>page 16</u>.

While there is little doubt that the sensation *color* cannot be defined, it may initially seem as if we are returning to the area of the definable, if our perceptual image is *colorless*: *gray values* are definable, aren't they? – Yes, they are, but the *sensation* associated with them *is not*: only the intensity of the light can be defined, and also the neuronal excitation that results from it. But when we move on to *perception*, we leave the realm of information: the *brightness* that we perceive is just as much a *sensation* as the *color.*

And the same applies to all other senses: the frequency of a sound can be defined, but the sound-*sensation* can not, etc.

This means:

If sensation is lacking, then there is no "inner theater", which is <u>made up of sensations</u>.

So to put it very clearly:

AI systems do not see, do not hear, do not feel, do not smell, do not taste.[13]

Unfortunately, our language is not suitable for distinguishing between system states *with* sensation and those *without*. For us, "seeing" or "hearing" simply means what it **is** for us, and that is in any case information **and** sensation. Therefore, *strictly speaking*, statements about perceptions are only correct if they refer to humans or higher animals, otherwise they are wrong: robots *do not see*, bees *do not see* – they only process frequency, intensity, distance and direction information.

B) The second area is the area of *feelings and moods*. Nothing further needs to be explained here.

AI systems experience nothing and feel nothing. They feel neither happiness nor unhappiness, neither love nor hate. They are neither cheerful nor sad, neither in a good mood nor irritated.

This list can be continued at will, since every mental state is a *quale*, i.e. consists not only of *information* but also of *sensation*.[14]

C) We have determined *sensation* as *substance of the mental state*. It follows that it must be understood as *cause of the mental dynamics*.

Accordingly, ***everything*** that drives our thinking and acting must have a component of *sensation*. There is no acting or thinking without a motive. Even purely logical reasoning can only take place if we ***want*** to find the correct solution.

Therefore also applies:

AI systems cannot want or not-want anything. They know neither motive nor interest, neither curiosity nor rejection.

In this area, the lack of differentiation in language use is particularly problematic. Programmers speak of the "goals" or "intentions" of an AI system, of what it "strives for". In all cases, however,

---

13 This also applies to simple animals, such as insects, for the following reason: We have shown that the emergence of sensation can only occur if the neural network develops its own, *internal* laws. A necessary (and sufficient) condition for this, however, is that the network contains *functionally unbound structures*, i.e. structures whose function is not determined genetically or by early programming. Only under this condition can (and will) the *network of neural states* (attractors) develop that we understand as *mind*.

For us, *having eyes* is synonymous with the *ability to see*. But this is wrong. For an animal that has a light-sensitive cell, the world is by no means *bright* – the animal only has the *information* about which direction the light is coming from.

14 Of course, there are also activities *without* sensation, such as reflex actions or automatically executed sequences of movements. However, these are not *mental* activities, but *neuronal* activities.

this is only an increase in a parameter value, and not *goals* or *intentions* as we understand them as elements of human action, which are always linked to emotions.

After this brief preparation, let us now turn to the question:

What does the absence of sensation mean for the performance of AI systems?

3.3. What can be ruled out with certainty

Alan Turing, from a 1951 lecture:

"It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. At some stage therefore we should have to expect the machines to take control."

Geoffrey Hinton, October 27, 2023, University of Toronto:

"Suppose you have multiple different super-intelligences. ... you're gonna get evolution of super-intelligences. And let's suppose there's a lot of benign super-intelligences who are all out there just to help people. ... But let's suppose that one of them just has a very, very slight tendency to want to be a little bit better than the other ones, just a little bit better. You're gonna get an evolutionary race, and I don't think that's gonna be good for us. So I wish I was wrong about this. ... My guess is that they will take over, they'll keep us around to keep the power stations running, but not for long. ... That's my best guess, and I hope I'm wrong."

Geoffrey Hinton, February 19, 2024, Oxford's annual Romanes Lecture at the Sheldonian Theatre:

"… what happens if super-intelligences compete with each other? ... As soon as they get any sense of self-preservation, then you'll get evolution occurring. ... the more aggressive ones will win. And then you get all the problems that Chimpanzees like us have: lots of aggression and competition." [15]

These quotes express the expectation shared by many AI experts of what could happen – or probably *will* happen – when AI becomes AGI (Artificial General Intelligence), that is, when it not only achieves or surpasses human performance in *certain* areas, but in *all*.

As a result, there would be an exponential increase in performance through self-optimization, so that AI systems, due to their vastly superior intelligence, would replace us as the dominant species. Just as great apes are currently at *our* mercy, in the future *we ourselves* would be dependent on the goodwill and mercy of AI systems.

Fortunately for us, as we know, Geoffrey Hinton's wish ("I wish I was wrong about this") has already come true before he even uttered it:

According to our proof that AI systems are not sentient, we will by no means create a species superior to us, but only ***emotionless, mindless zombies*** – mere ***automatons*** that are ***not even capable of perceiving*** anything.

So they will not "take control" under any circumstances, because they cannot ***want*** to, they will neither like nor tolerate us, neither despise nor destroy us, in fact it would even be inappropriate to claim that we are *indifferent* to them – there is simply ***nothing*** there.

---

15 It was these statements – and many others like them – that prompted me to add another part to my work on free will and artificial intelligence. I would have found it difficult to leave such fundamentally false claims about our future unchallenged.

In summary:

AI systems are not a new, super-intelligent, dominant species. They are not living beings, but automatons.[16]

But AI systems without sentience and consciousness can also pose dangers.

Let's hear Stuart Russell, December 14, 2023, Penguin Channel:

"… that's the key in the Skynet story: It becomes self-aware. That's a very common idea in science fiction, both in books and particularly in film. In AI-related films there's gotta be a struggle between AI and humanity. And the way that struggle happens is almost always because the AI becomes conscious. We call it the Spooky Emergent Consciousness meme. It's really a red herring. Because of its frequent occurrence in film, one often sees this in serious journalism as well, but in fact we need to worry about machines not because they're conscious, but because they're competent. They may take preemptive action to ensure that they can achieve the objective that we gave them. That's the real concern. So if someone tells you, 'Don't worry: as long as it doesn't become conscious, everything's fine', don't be reassured."

Except for the expression "preemptive action", which implies planning and intention, Stuart Russell's claim is not called into question by our proof against AI sentience – at least not directly. In the next section, however, we will show that the proof contains arguments that set limits on the possible competence of future AI systems.

3.4. Which limitations are likely

Given our proof that AI systems are not sentient, it is self-evident that the popular dystopia in which we, an inferior species, are at the mercy of an unimaginably powerful superintelligence, is simply nonsense.

But now we must grapple with the far more difficult question of how the lack of sentience limits the performance of AI systems.

First, we can state the following:

There is a striking connection between the fact known as ***"Moravec's Paradox"*** and the fact that ***AI systems cannot perceive***.

In 1988, Hans Moravec wrote:

"It is comparatively easy to get computers to perform at adult levels on intelligence tests or at playing checkers, and difficult or impossible to get them to perform at the level of a one-year-old in terms of perception and mobility."[17]

---

16 In living beings – which owe their existence to biological evolution – insensitivity is linked to the condition that the neural network is simple and has only a very limited ability to learn. To follow up on the deliberations in the second part: a system is only *alive* if its **activity** arises from the *inseparable unity of substance and accidents*, that is: if it is *essential* and *happens by itself*.

From this follows that life – just like mind – *cannot be constructed*.

17 Moravec explained this surprising fact by arguing that evolution had much more time to perfect our sensorimotor skills than to develop our logical-abstract abilities.

I think this argument is inadequate: complex movement sequences must first be developed in the motor cortex – a structure of the neocortex, the most recent part of our brain in evolutionary history – and only then are they stored in the much older cerebellum. The skills in question are therefore just as new as the ability to think logically, which also takes place in the neocortex. (Try teaching an orangutan archery – it

(To update the paradox, replace "checkers" with "Go" and the "one-year-old" with a "five-year-old.")

In the previous sections, we established that there is no *"inner theater"* in robots: **they see nothing** – in the sense that they have no *"inner image"* of the surroundings.

It is therefore natural to assume that the sensorimotor difficulties of AI are due to its inability to *perceive* the environment in the way that is so self-evident to us.

Why is that? Why shouldn't *information without sensation* be just as suitable as basis for manipulating objects and finding our way in the world as *our vision* is?

The intuitive answer is clear and unambiguous:

Intuitively, it is evident that the ***"image"*** of the environment that is always present to us ***as a whole*** – not only sensorially, but also in its *meaning*, including all the objects contained therein and their relationships – is superior to the information given pixel by pixel to an almost fantastic extent: the information must first be assembled and then analyzed, numerous recognition processes must be carried out, the possible relationships of the recognized objects must be determined in terms of their suitability to be part of the overall scenario, the meaning of which must also first be determined, and so on and so forth.

With equal certainty, also the following can be stated:

In order to *understand* what we see, we need precisely this "image" just outlined. But it seems to be generally true that the mental state that we call ***understanding*** presupposes a scenario of exactly the same kind as this ***perceived*** image: an ***imagined*** "image" in which the objects and facts are gathered that we need to understand the overall situation.

The ontological-analytical view supports this claim:

We have shown that *mind* is the *causal level* of the neural network. We therefore do not understand mental activity as the *dynamics of neurons*, but as the *dynamics of mental states*, i.e. of the neural patterns that we have determined as *attractors* of the neural dynamics. *Mind* is therefore to be understood as a *network of attractors*.

This means that, unlike with pure information without sensation, we do not need to recognize, analyze, put together, relate, estimate possible consequences, etc.

Why? – Remember: an important property of attractors is that the system only needs a small subset of the set of parameter values that are within the catchment area of the attractor to create the attractor state.

Here is a simple example:

For us, observing *short red pants* and a *small rolling object* can be enough to conjure up the mental image of a "child chasing a ball", including the possible consequences – i.e. information that can be extremely important while driving a car.

---

will be just as unsuccessful as trying to optimize its logical abilities.) Conversely, we probably won't be able to do math particularly well for many millions of years. (If we still exist then.)

[I found the following funny version of Moravec's Paradox on the Internet:
"I used to think that at some point in the future I would finally have enough time to write poetry and paint while my robot tidied and cleaned. But things have turned out completely differently: Now I have plenty of time to tidy and clean while my robot writes poetry and paints."]

"Mental states" are therefore always *wholes*, just like the sensory states we have just talked about: the "image" of the environment or the "inner mental image", which are always given to us *as a whole*. They already contain all the details and connections that must first be recognized and analyzed individually in pure information processing.

It seems that sensation, the indefinable element of our mind, is responsible for this integrative performance – the presentation of the whole picture including the connections of all the details – which is precisely what AI systems lack.

However, the following must be taken into account:

Since artificial neural networks remain insentient even when they develop attractors, (see Part 2, page 22, first note), the attractor argument is suitable for illustrating the integrative performance of sensation, but the existence of attractors can only be a necessary and by no means a sufficient condition for it.

The *actual* reason for this performance lies in the following:

In the case of a *really existing* object, there must be something from which the *activity* of that object emanates. We have called this element of its existence *substance*. The first step in our proof that AI systems are insentient (Part 2, Section 2.2) was to show that ***sensation is the substance of mental states***. So ***sensation is what drives the mental activity***.

Thus, in a neural network that produces mind, *sensation* has the same status as *mass* in a system whose dynamics is caused by gravity, such as our solar system. Just as mass determines the interaction of the objects of the solar system, sensation determines the interaction of the objects that make up our mind, i.e. of our mental states.

This means:

Just as <u>mass</u> guides the objects of a gravitational system and <u>connects them with each other</u> – like e.g. Earth and Moon – so <u>sensation</u> guides the objects of a mental system and <u>connects them with each other</u> – like e.g. child and ball – and it does so, like mass, <u>by itself</u>.

Can this connection be imitated in a simulation? Not in every case, since there is an absolute boundary between original and simulation. As follows:

We know: The dynamics of a system ***without gravity*** can never completely match the dynamics of a system ***controlled by gravity***.

Therefore it must also apply: The dynamics of a system ***without sensation*** can never completely match the dynamics of a system ***controlled by sensation***.

So there is an absolute limit between artificial intelligence and the human mind.

However, we do not know where this limit is.

With gravity, we have a mathematical description of the interaction, so we can at least estimate the limits of a simulation.

With sensation, this is not possible: in this case, the attempt at a mathematical description is completely futile. How could one succeed in describing the dynamics of a system that consists of objects – attractors – whose more complex forms are already difficult to control mathematically in and of themselves, and which, moreover, are *constantly changing* as a result of their interactions?

Here, the non-computability of reality reaches a level that certainly excludes mathematical access – not only now, but also in any conceivable future.

We are therefore, on the one hand, dependent on ontological arguments, and, on the other hand, on what self-observation tells us about our own thinking – how we recognize, generalize, explain, conclude, etc. – and what follows from this.

But I want to stop here and continue the discussion later, and start another line of argument that relates to currently available types of AI systems.

Fortunately, as of late we are able to assess both the capabilities and limitations of current AI.

This has long been the case for symbolic AI – the "classic" way of programming AI systems, where a logical structure is built from defined elements – but for self-learning neural networks – such as Generative Pre-trained Transformers (GPTs) – we have only recently learned what incredible feats they are capable of and yet what strange limitations they are subject to.

To demonstrate this and prepare the general assessment of AI's capabilities, let us first look at some instructive examples.

We start with a simple neural network that we want to train to recognize handwritten digits.[18] To do this, we need a sufficiently large set of pictures of such digits.

A *training run* consists in presenting all the elements of this set to the network.

Input is the gray values of the pixels of these pictures.

First, we assign random numbers (called *biases*) to all neurons (except those of the input layer), which – in analogy to biological neural networks – represent the initial activations of the neurons.

We also assign random numbers (called *weights*) to the connections that lead from all neurons in one layer to all neurons in the next layer, which express the *strength of the influence* of a neuron on the neuron connected to it.

Since the initial values are random, on the first run the recognition rate will not be higher than that of a random generator.

Then we try to improve the performance of the system by changing the weights and biases before the start of each subsequent run. We therefore consider them as *variables*.

Our goal is to minimize the *error rate*. It can be viewed as a *function of these variables*. So we are looking for the *minima* of this function.

In this way, with relatively simple means and after a large number of runs, it is possible to achieve a recognition rate close to 100% not only with the training set, but also with any other set of handwritten digits.

Although the neural network is simple and the task assigned to it is limited, it gives rise to precisely the questions and hypotheses that we want to investigate below.

The first question is:

According to which criteria does the network recognize the digits?

---

18 At this point I had actually planned to give a description of the network, which would also serve as a short introduction. However, I soon abandoned this plan: for people who are completely unfamiliar with neural networks, the introduction would have been too short and therefore not helpful, and for everyone else it is superfluous anyway. Instead, I refer to the page https://www.youtube.com/@3blue1brown, where an excellent multi-part introduction is available under the keyword "Neural Networks", which is also very nicely presented graphically.

In my own presentation, I will limit myself to the facts that are important for my later argument.

In any case, we know how *we ourselves* proceed:

We see each digit as composed of clearly defined components, e.g. the 3 made up of two semicircles open to the left and placed one above the other, or the 4 made up of three sections of straight lines arranged in a certain way, etc.

Our way of recognizing the digits therefore arises from the *construction* of the digits. **We know the digits** and perceive them as being built up piece by piece.

Does the artificial network proceed in the same way? That is extremely unlikely, since **the network does <u>not</u> know the digits**.

This may sound strange, since it is capable of *recognizing* them. But this recognition does not occur, as with us, by *comparison* with the *mental image* of the number.

In the network, the recognition process is based on a completely different principle: in fact, the numbers are *not directly represented* in the network, or let's say: only *implicitly* and not *explicitly* represented. A comparison is therefore not possible.

But through training, the network has found one of the (local) minima of the above-mentioned function, and has proceeded according to criteria that are completely incomprehensible to us.

The two methods are mutually exclusive. It must therefore be assumed that our method of recognition does not correspond to a minimum of the function in the network's search space.

It is astonishing that there are any other ways of finding criteria for digit recognition apart from our method of shape analysis. In any case, *we* are not able to imagine such a possibility. But this is undoubtedly because the function whose minima are sought is located in an extremely high-dimensional space – the number of its dimensions (which is equal to the number of the variables) can be greater than 100,000 even in relatively small and simple networks, while our imagination is limited to spaces with a maximum of 3 dimensions.

In comparison with the number of dimensions of the *search space*, the number of dimensions of the *recognition space*, in which the network actually identifies the digits after the training has ended, is relatively small: the *weights* and the (initial) *biases* are now constant, the only variable quantities are the *activations* of the neurons in the hidden layers (the neuron layers between the input layer and the output layer), which result from the respective input. They are therefore the quantities that can be understood as coordinates of the recognition space.

So the number of dimensions is equal to the number of all neurons minus the number of input and output neurons.

To each digit a subset of this recognition space is assigned – precisely that subset into which all inputs lead that come from pictures in which the network activates the output neuron associated with this digit (however, these can also be nonsense pictures or random pictures).

The union of all these subsets is the subset of the recognition space that consists of *all* values of the activations that can occur as a result of every possible input.[19]

---

19 This unusual view serves to define the terms "implicit" and "explicit" more clearly:

We are not considering the *function* that the network applies to the input, but the *space* in which the recognition process takes place. This is the best way to understand what is meant by the fact that the digits are only "implicitly" represented in the system and not "explicitly": the subset assigned to the digit 2 can certainly not be considered an "explicit" representation of the digit 2, and if – given a picture of the digit 2 as input – the activations of the network neurons take on values that correspond to the coordinates of a point located in the subset assigned to this digit, this does not mean that the network **knows** the shape of the digit 2 – not in any possible sense of the word "know".

Despite this strong reduction, the number of dimensions is still far too high for our imagination, and even mathematical analyses do not give us any understanding of the criteria by which the network recognizes the numbers.

This *could* be interpreted as an indication of the enormous possibilities of AI systems to recognize connections – regularities or semantic structures – in a way that is superior to ours to an unimaginable extent – or is it? We will come back to this.

The next scenario we will consider is the Go competition from 2016 between the neural network AlphaGo developed by Google DeepMind and the South Korean Go master Lee Sidol, who many experts considered the best player in the world at the time.

Before this competition, Go was considered the domain of human intelligence and creativity because – due to the enormous number of possible courses of the game – it was still inaccessible to *symbolic AI*, which had already left humans far behind in chess, and because self-learning neural networks were still hardly known and tested.

AlphaGo won 4:1. Its 37th move from the second game became legendary: it was a move that was considered forbidden according to Go theory, which had been constantly evolving for centuries, because it was believed that it would result in positional disadvantages. As it turned out, however, it was the winning move.

AlphaGo was celebrated, its creativity was called "unique", and the program was awarded ninth dan, the highest rank possible in Go, on the grounds that its game had almost reached "divine regions".

It should also be added that, a short time later, AlphaGo – which still had an extensive database, was beaten 100:0 by AlphaZero – which had no knowledge of Go other than the Go rules and had improved its game solely through optimization in billions of training runs against AlphaGo.

So it seemed to be proven that human intelligence is hopelessly inferior to artificial intelligence in the form of self-learning neural networks, not only in terms of logical thinking, but also in the area to which we believed to have exclusive access: the area of creativity.

Is our fate thus sealed? Not at all! – The story is not over yet:

At the beginning of 2023, an AI research team reported that it had succeeded in developing a strategy that could beat the best Go programs.[20]

Kellin Pelrine, a member of this team and a good amateur level Go player, beat KataGo – one of the strongest neural networks – 14:1.

As Stuart Russel reported,[21] Pelrine beat KataGo even when he gave him 9 stones in advance, and in this case even 15:0. He also beat other equally powerful Go programs developed by different teams using different methods.

How is this possible? There is a clear answer to this:

The programs have ***no idea about the game principle*** – they ***do not know*** that it is about enclosing regions and opponent's stones. As can be seen from their lost games, they ***do not realize*** that they are being enclosed and therefore allow it to happen, even though they have several moves to prevent it.

---

20 https://arxiv.org/abs/2211.00241

21 Stuart Russell, "AI: What If We Succeed?" April 25, 2024, Institute for the Study of Ancient Cultures Museum.

Why do they still show such incredible playing strength in games against Go masters?

Because – just like the digit recognition program – they have optimized their performance by looking for and finding minima of a function of an enormous number of variables in an extremely high-dimensional space. In doing so, they have discovered game strategies that we could never think of, but on the other hand, they have no chance at all of defending themselves against the human strategy – in this case Kellin Pelrine's – because it is not close to a minimum in their high-dimensional search space.

Why is that?

The answer is very similar to that of the previous example:

When discussing digit recognition, we found that the network *does not know* the digits, and indeed *cannot* know them, and that therefore our recognition method, which is constructive and based on comparison with the imagined digit, does not correspond to a minimum of the function in the search space.

When playing Go, however, we cannot speak of *our* – i.e. the *human* – strategy in general: people have very different strategies in playing Go.

But we can state the following:

*The Go game principle is the most important prerequisite for **all** human strategies.*

On the other hand, we know:

The neural network ***does not know*** the game principle.

Nevertheless, the game principle must be *implicitly* present in the AI system and play a role in its strategy – otherwise self-learning neural networks would be completely incapable of playing Go. The systematic optimization of the game strategy *can* only succeed if the game principle is involved in this process.

But that means exactly this:

*Although the principle is **a prerequisite for the optimization process**, it is **never integrated into the system** itself.*

So here, too, it is the same as with number recognition, where the shape of the numbers must also be *implicitly* present in the search process, but does not exist *explicitly* in the system: although the system recognizes the numbers, *it does not know them*, and in the same way applies: although the Go program uses the game principle almost perfectly in its victories, *it knows nothing about this principle*.

So what is the relationship between human and AI-generated playing style?

This can be seen from the results available:

Obviously, human Go masters *always* stay in game scenarios that are located in the "valleys" of the high-dimensional function that the AI systems have explored on their way to the local minima. In this case, humans have no chance because the AI systems are *in any case* closer to the minimum.

It follows that humans must stay as far away as possible from the areas (game situations) that the AI systems have explored during their self-optimization. Simply put: They must not play too well.

The much more important second condition is that they must concentrate on what AI systems *do not recognize*: on encircling the opponents' pieces – in the sense of the first rule especially when the moves required for this are actually *bad moves* because they do not serve the build-up of the game.

As it turns out, humans have excellent chances of winning if they follow these two tactical instructions.

In short:

Neural networks that achieve their playing strength through self-optimization have *no chance of understanding the game principle*. Humans who are able to exploit this limitation are clearly superior to them.

Unfortunately, in 2016 neither Lee Sidol nor anyone else knew about this fundamental weakness. Otherwise Lee Sidol would have won easily, and the assessment of AlphaGo's performance would certainly have been very different.

Earlier – in [section 3.2](#) – we showed that the proof that AI systems are not sentient forces us to use language in a more differentiated way, to define several words from the area of perception and motivation more precisely.

The same thing now happens in the area of performance assessment. With humans, it is self-evident that astonishing performances, such as the 37th move from the 2nd game, can only come from a *deep understanding* of the game. So calling them *creative* has so far been firmly linked to this fact. You can still use this term, but when applied to the performance of a neural network, it is *redefined*, since this performance is no longer *due to deep insight*, but quite the opposite, *despite its complete lack*.

Part of the meaning of the word *creative* would remain, however, since something *new* was actually discovered. But if I asked you directly: "Would you call an unbelievably speedy idiot who runs downhill until he stumbles upon something that is so deep down that no one has found it before *creative*?", you might hesitate.

The change that the term *understanding* would undergo if it were applied to AI systems would be even more dramatic:

Since *understanding* has been a self-evident prerequisite for great intellectual performance before the development of AI systems, the person who achieved such a performance had to be recognized as having understanding in any case. But if a neural network achieves the same – or even a more significant – performance, then the attribution of understanding would *completely rob* this term of its meaning – its use would simply be grossly wrong.

Can this weakness of AI systems be corrected?

Let's hear Stuart Russell on self-learning neural networks in general:[22]

"...if you have a very, very large representation of what is fundamentally actually a simple concept, then you would need an enormous number of examples to learn that concept. Far more than you would need if you had a more expressive way of representing the concept." –

– where "more expressive way" means a programming language such as *Python*, in which, for example, the Go game principle can be expressed very easily.

However, Stuart Russell describes the problem very cautiously here, because in fact – if the task of the AI system concerns a *real* and therefore not fully definable scenario – an *infinite* number of examples would be needed to fully integrate the concept into the system.

Although the Go game consists of a finite number of definable states, still their number is so large that it is not possible to exclude all successful counter-strategies.

In other words: The Go game principle *cannot* be fully integrated into the AI system.

---

22 Stuart Russell, l. c.

At this point, the question arises whether this deficiency of self-learning neural networks could not be remedied by supplementing them with symbolic AI.

We will deal with this question below. But now we turn to our next example, the type of AI system called *Generative Pre-trained Transformer.*

What is a GPT? What can it do? – I will outline the answer here only to the extent necessary to be able to follow on from the previous considerations.

GPTs are learning neural networks that are able to emulate large systems of *structured* data and produce something on this basis – texts, images, translations, etc.

In the case of LLMs (Large Language Models), this means:

They capture the grammatic, syntactic and semantic structure of the language in total, and also the semantic structures of specific linguistic creations, not only sentences, but also larger units – stories or literary works of various kinds – in other words: they are also able to emulate context-dependent structures.

Achieving this performance requires an immensely complex training phase. The learning process is prepared by breaking the data down into small elements, so-called tokens – in the case of language this means: words, parts of words, syllables or even letters, in the case of images: motifs, image sections or pixels; in the case of acoustic data: characteristic elements such as tones or noises, or simply short temporal sections, etc.

(For the sake of comprehensibility, in the following I will limit myself to linguistic tokens.)

First, a list of all tokens that occur in the data set is created.

To the tokens Vectors are assigned. (In GPT3, these vectors have more than 12,000 components.)

The tokens are therefore represented by vectors in a high-dimensional, abstract space.

The numerical values of the components are initially random (as in our two previous examples).

The learning process consists in the GPT being presented with sections of text. Its task is to find the *next word*, i.e. the word that follows the respective section.

It is clear that it can only succeed if its vectorial representation of all tokens – and thus of all words – reproduces the grammatical and syntactical structure of the language in general, and in particular the semantic structure of the text in question.

The error rate can be understood as function of the vector components. They are therefore the variables of this function, and the aim here is again to find minima of the function.

One can anticipate that achieving this goal is only possible with enormous effort: the amount of data is huge – basically all sentences available on the Internet – and the semantic structures of language products are complex and ambiguous. This is why previous attempts with self-learning neural networks failed. Only the extremely increased storage and computing capacity has made the current success possible.

To put it simply, the training process is an investigation of the degree of "closeness" or "cohesiveness" of words, and also their "relatedness". At the end of this process, words with similar meanings should be represented by vectors pointing in similar directions.

One effect of the representation by vectors is that *directions* in this high-dimensional representation space are *elements of the semantic structure*. A well-known example of this is that the vector (woman minus man) corresponds almost exactly to the vector (aunt minus uncle), or the vector

(daughter minus son). The three difference vectors are almost parallel and of the same length, and their direction means "(change of) gender".

In parallel to the semantic contexts, the grammatic and syntactic rules of the language must also be learned: word classes, sentence structure, etc.

I want to stop here, because despite the incompleteness and anecdotal nature of this introductory sketch, what has been said so far is already sufficient as background for the questions that we now want to ask again:

We already know how the network represents the words. What we don't know, however, is the criteria according to which this representation takes place.

We know how *we ourselves* proceed: we also have such a "representation space", even if we are not directly aware of it. We define words by *properties*, and thus these properties are our criteria: the components of our vector representation and therefore also the coordinates of our representation space.

Does the network proceed in the same way? Certainly not. Its representation space is completely abstract, and the coordinates from which it is constructed actually have no concrete meaning at all. There could be any number of them – the more the better, if the amount of data is large enough and the computing power is sufficient. Of course, due to the structural similarity of the two spaces, there must be a statistically researchable connection between the GPT *criteria* and our *properties*, but there is nothing more to say about that. The components of the vectors of the GPT representation are abstract and meaningless.

I propose a thought experiment that is an extension of Ronald Searle's "Chinese Room":

The GPT receives the data of the complete speech production of an alien civilization – just as it previously received the data of the earth's speech production.

It now carries out the same kind of training.

After that, you can chat with the aliens – you just have to present their messages to the GPT and then reply to them with what the GPT has produced. (You could also do the GPT's calculations yourself, but the duration of the universe's existence would hardly be long enough for that.)

So you have a great conversation, tell each other jokes, and become good friends. Or do you?

Well, friendship is unlikely to happen. You have no idea what you were actually talking about. Maybe it was about the aliens' favorite pastime, eating members of other civilizations?

But wait! – maybe the GPT understands something about your communication?

No. Just as the digit recognition program does not know the digits, the GPT also knows nothing about the meaning of words – its performance is based on the statistical data of their occurrence, which result from the (grammatic, syntactic and semantic) structures of language production, which in turn follow from the statistics.[23]

---

23 This is also the reason why Searle's argument is now inadequate. Searle himself has repeatedly emphasized that the execution of a program only requires knowledge of a sufficiently large number of rules, and he believes that from this follows that the *semantic structure* (which he equates with understanding) remains excluded. He failed to notice that GPTs have exceeded this limit, precisely because the *statistical relationships* also **contain** *a large part of the semantic structure*.

However, the core of Searle's argument remains untouched: Just as it is impossible to conclude *understanding* from correct language production based on a fixed set of rules, it is also impossible to conclude it from a correct language production based on the probability of the occurrence of words. In this way, Searle's argument can be transferred from symbolic AI to neural networks.

But here again the same applies as before:

*Although the grammatic, syntactic and semantic structures have **guided the optimization process**, still **none of them is explicitly present in the GPT**.*

So it knows as little as you do.

In other words, ***the GPT understands*** as much about the communication with the aliens as it does about the communication with humans, and that is ***precisely nothing***.

Here we encounter the same language problem as in our previous examples:

If *humans* talk sensibly, then it would be absurd to claim that they do not understand what they are saying.[24] But now we are forced to abandon this firm connection between "talking sensibly" and "understanding". Just as neural networks can correctly recognize something (recognize in the sense of identifying it) without knowing it, and just as they can play Go brilliantly without even knowing what it is about, they can also *talk* sensibly without *being* sensible and without understanding anything about it.

To complement this, I will now give a few more examples. I am taking them from Yejin Choi, a computer scientist and professor at the University of Washington.

Yejin Choi, 28.04.2023: Why AI Is Incredibly Smart and Shockingly Stupid (TED Talks #ai):

" … suppose I left five clothes to dry out in the sun, and it took them five hours to dry completely. How long would it take to dry 30 clothes?

GPT-4, the newest, greatest AI system says: 30 hours. – Not good.

A different one: I have 12-liter jug and six-liter jug, and I want to measure six liters. How do I do it? – Just use the six liter jug, right?

GPT-4 spits out some very elaborate nonsense:

Step one, fill the six-liter jug.

Step two, pour the water from six to 12-liter jug.

Step three, fill the six-liter jug again.

Step four, very carefully, pour the water from six to 12-liter jug.

And finally you have six liters of water in the six-liter jug that should be empty by now.

OK, one more.

Would I get a flat tire by bicycling over a bridge that is suspended over nails, screws and broken glass?

Yes, highly likely, GPT-4 says, presumably because it cannot correctly reason that if a bridge is suspended over the broken nails and broken glass, then the surface of the bridge doesn't touch the sharp objects directly.

OK, so how would you feel about an AI lawyer that aced the bar exam yet randomly fails at such basic common sense?

AI today is unbelievably intelligent and then shockingly stupid!"

---

24 Except for trivial cases, such as reciting a text from memory or reading it.

While I think arguments are much more important than examples, it is still pretty impressive how these three examples show that the GPT *doesn't understand at all what it's actually about*. The first two examples in particular show that it produces complete nonsense when there are not sufficiently similar scenarios in its training data – or when it simply chooses the wrong ones because it does not capture the causal relationships.

It is also clear that these cannot be just minor "glitches": *humans* may be susceptible to such glitches – even if they are intelligent.

But if *AI systems* are intelligent, they are either ***always*** intelligent or ***never***. So the wrong answers of the GPT must be taken as ***evidence of complete lack of intelligence***. Whether the answer is right or wrong is simply a matter of *chance* and not a question of intelligence.

For this reason, I am surprised that even critics of the current AI euphoria such as Yejin Choi or Gary Marcus express their reservations so cautiously: they speak of "temporary deficiencies" or even of "strategies for correcting them", although the principle behind the failure of AI is clearly recognizable and actually ***cannot be corrected***.

What is this "principle"?

Exactly the one we encountered in our three examples:

In neural networks, *learning* means the following:

Their performance is optimized by searching for the minima of a (high-dimensional) function in numerous training runs, the variables of which correspond to the – initially random – weights and biases of the neurons.

The *value* of this function (the "error rate") *can only decrease if the formal and structural conditions of the desired performance control the search process*.

In our examples, these were:

> – *the shape of the numbers,*
>
> – *the basic principle of the game of Go,*
>
> – *the semantic structure of the previous word string.*

In these three cases – and in all other cases – these conditions indeed control the optimization process, but they ***remain*** merely ***prerequisites of the AI system*** and ***never become part of the system itself***.

In other words:

The system knows nothing about them, it does not recognize them, it does not understand what is the issue – or whatever you want to call this fact.

Can this fundamental deficiency be remedied?

Basically, there are two ways to reduce the number of errors, i.e. of nonsense or undesirable behavior:

> 1. You can influence the training process using previously defined rules.
>
> 2. You can control the output using a catalog of instructions.

The first method can also be carried out by humans. The second method means supplementing the self-learning system with symbolic AI.

However, to both types of improvement the well-known law applies – which also generally limits the performance of symbolic AI:

Every catalog of rules necessarily remains incomplete because in the real world constantly new situations arise.

We are now experiencing numerous cases of this fact: the development teams are eagerly trying to correct the stupidities of the GPTs, and the critics are finding ways to make the developers' corrections ineffective again with small changes, or they are looking for new errors.

So it is not at all about clarifying the question:

"Can the lack of understanding be fundamentally eliminated?"

– this question has long been answered, and the answer is ***no*** –

but about the question:

"Is the respective catalog of corrections sufficient for the intended purpose?"

The crucial point is that symbolic AI is in no way suitable for eliminating the complete lack of understanding. AI systems in the form of self-learning neural networks *do not understand anything*, and (of course) the addition of symbolic AI cannot change this – it can only reduce the number of errors.

The most important consequence of this fact is that the currently prevailing AI technology is unsuitable for producing AGI.

AGI is based on generalization. However, *understanding* is a necessary condition for all types of generalization.[25] In order to transfer the causal structure of one process to another process, for example, it is necessary to *understand* this structure – all three examples from Yejin Choi show this very clearly.

Here again, one can try to create a catalog of transferable causal relationships, but this catalog will remain *extremely* incomplete.

What about future AI? Or more specifically:

Can the limitations of current AI be overcome by future hardware and software?

So this question concludes my two strands of argument, and what can be said about it will be the subject of the final section.

---

25 This also applies to trivial types of generalization: e.g. any form of compression with loss is a generalization, since leaving details behind corresponds to progressing to the general. But here, too, understanding is required, because it must be known which properties the causal structure depends on – otherwise the generalization is a matter of luck (as with the GPT).

3.5. Overview, comparison, final assessment

The starting point of my argument about the limits of artificial intelligence is the proof presented in Part 2 that AI systems have *no sensations*.[26]

This means:

1. *AI systems cannot <u>perceive</u> anything*.
   They lack the *"inner theater"*, the *"image"* of the environment: they cannot *see*.
   Likewise applies: they cannot *hear*, *feel*, *smell* or *taste*. For them there is only *information*.

2. *AI systems cannot <u>experience</u> anything*.
   They have no feelings.

3. *AI systems cannot <u>want</u> anything*.
   They lack intentionality and motivation.

In section 3.3, we pointed out an obvious consequence of this proof:

No matter what the future of AI may look like, due to the limitations mentioned above AI systems will *never* be a new, superior species. The dystopias in which we are at their mercy belong in the realm of fantasy.[27]

In section 3.4, we linked the *lack of perception* in AI systems to *Moravec's paradox*, and then outlined the amazing integrative feat that our perception achieves. Finally, we recalled the existence of an absolute boundary between a system and its simulation, which was proven in part 2, and thus also exists between human mind and artificial intelligence.

These are all indications that the ability to perceive gives us a considerable advantage compared to systems without perception. However, for the moment it remains open to what extent the disadvantage of the lack of sensation of AI systems can be compensated for by the increase in storage capacity and computing power as well as by the further development of the system architecture.

We then turned to a different argumentation strategy: examining the limits of the currently prevailing AI technologies.

This investigation then led us to the following insight with surprising clarity:

Self-learning neural networks do not understand anything of what they produce.

They do not know what they recognize, they have no idea what they are talking about, they know nothing of the principles that enable their performance.

We have also found that Symbolic AI cannot remedy this fundamental deficiency. It can only reduce the number of errors caused by it.

Since understanding is a necessary condition for (meaningful) generalization, from this follows that the types of AI that are currently so popular cannot be further developed into AGI.

---

26 Here – as always in this paper – "sensation" stands for that part of a mental state that is *not definable*, that is, that goes *beyond information*.

27 E.g. this applies also to Nick Bostrom's popular "Paperclip-Scenario" – already the title of the respective paper:"The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents" is sufficient to refer the scenario to the realm of fantasy, since the agent has neither *will* nor *motivation* –, as well as to Marvin Minskys Riemann hypothesis catastrophe.

This raises the question again of whether these limitations can be overcome by improved hardware and software.

Before we turn to this question, we will carry out a brief comparison that will help to further shed light on the topics discussed so far: the comparison between the way *we* think and the way ***neural networks*** do it.

For that purpose, we consider a human train of thought in exactly the same way as we have done with artificial neural networks so far, i.e. not as a *train of thought* as we usually do: as a sequence of assumptions, guesses, conclusions, errors, planning, etc., but as a *neural process*.

Let's assume we make a move in a game of Go.

Input is the position of the game, output is our move. So this move is to be understood as the result of the function that our neural network performs on the input.

As in our example scenarios with self-learning neural networks, the excitation states of the neurons involved and their connection strengths are the variables of this function.[28]

At the end of the calculation process, our hand will make the move according to the output.

We now ask the same questions as in our examples:

The weights and biases of the neurons involved are the coordinates of an extremely high-dimensional space. They can be viewed as *criteria* or *components* of the decision that the neural network ultimately makes.

Do they have any meaning? Obviously not. One could claim – as in the examples – that there is a structural correspondence between *this space* and the *decision space* in which the *corresponding train of thought* takes place, but there is nothing more to say about it.

Does **the network** "know" *what* it is doing and *why* it is doing it? Certainly not in this view.

But **we ourselves** know what it is about, of course, and so we come to the conclusions to which this thought experiment leads us – or rather, to which we are forced by it:

1. In the case of a human player, the neural view ***lacks what <u>really</u> plans and carries out the move: the human mind***. In this case, however, it is only missing *in the description – in reality* it is present.

2. In the case of the AI system, however, the ***mind is not only missing in the description, but also in reality***, and therefore it is impossible for the system to understand anything.

At this point we encounter a fact of utmost importance – exactly the fact with which this work began and on which it is based:

Our mind can only accomplish this feat if the mind is the <u>causal level</u> of the neural network, and that in turn is only possible if ***physical causality is <u>incomplete</u>***.

If our mind were merely the execution of physical laws, then trains of thought would be without any meaning and every conclusion would be an illusion.

This is actually self-evident: the idea that thinking itself leads to correct results obviously *presupposes* its causal effect: how else would it be possible to correct an error? – If my *thinking itself* were not causal – would *physics* have been wrong?

---

28 Of course, the situation is much more complicated than in current AI systems. But that is irrelevant for our thought experiment.

Whenever I believe that I have asserted something **because** it is correct, I have *presupposed* the causality of my thinking: only under this presupposition can one thought **follow** from another thought.[29]

If *mind* is the causal level of the network, then it follows that the neural approach outlined above is not just **incomplete**, but even **wrong**: the proof that physical causality is incomplete excludes the existence of a function that produces the output from the input and whose variables are the weights and biases of the neurons.

There are several reasons for this – the simplest is that the neural system *changes* during the decision-making process. (Imagine a function f(x) = y in which the x-axis ripples unpredictably while the function is being performed.)

What we have just done is basically a repetition of the argument about free will, only this time we have included artificial neural networks.

I will briefly summarize again:

For ourselves, the existence of a *mental level* can be presupposed; what we had to show was that it is the *causal level* of the network (which we did in the first part).

In the case of artificial neural networks, our examples and their generalization have already made it clear that AI systems, whose output is the function of variables corresponding to the states of individual neurons and their connections, do not understand what they produce.

This leads to the need to move up to the level of *neural ensembles*. The comparison with human neural networks just made confirms this need.

In addition, the comparison also shows that this level must be *self-dependent*, in other words: its dynamics must not be the *logical consequence* of the neural layer. Only under this condition can it be considered the *causal level* of the network, and only then can we claim that the system is capable of *thinking*.

Can this condition be met on the basis of current hardware? It seems that here, as before, each state follows from the previous state. In such a logical structure, causality from below is complete, and therefore there can be no self-dependent level of neural ensembles with their own dynamics above the neural layer.

That would mean:

There is no thinking and understanding in AI systems that run on such hardware – even if these systems are suitable for forming attractor networks.

But since constructed AI systems are not sentient in any case (see part 2, page 19 below), and because we will now show under this assumption that they are not capable of understanding anything *new* on any type of hardware, we can refrain from answering the difficult-to-answer question about the possibilities of future hardware.

This will therefore be the last step in our argumentative path: In order to determine how far the consequences of the proof of the insensitivity of AI systems actually extend, we will now assume

---

29 I consider it a *grotesque* of our intellectual history that this fact is not taken into account – indeed not even noticed – by philosophy, natural science or AI research. From the first French materialists in the 18th century to the present day, physicalists and determinists have doubted the existence of *morality*, but *thought* always retains its independence – otherwise they would not be able to *argue* at all – although it is quite obvious that it dissolves into physics just as morality does, if it is not *itself* understood as causal.

One has to decide: causality lies **either** in thought **or** in physics – both at the same time are not possible.

that hardware and software are no longer subject to any restrictions, that everything that is physically possible is also feasible.

The difference between the constructed and the natural system will still be that the activity of the natural system is *essential* – that is, it follows from the *inseparable unity of substance and accidents* and thus unfolds *by itself* – while the *constructed system* depends on *supplied* activity.

According to our proof, this means that the biological system is sentient and the artificial system is not (see part 2, center).

This leaves only one question:

What does the lack of sentience mean?

We begin by considering an LLM. We have found that LLMs do not understand what they are talking about.

We have determined that the reason why they nevertheless *appear* to be reasonable is the following:

During the training phase, the *semantic structure* of the language is – via the statistics of the distribution of words – increasingly *incorporated* in the probability calculation of the next word. This structure is therefore involved in *controlling the learning process*, but it is *never integrated into the system itself* – the system *does not know it*.

This abstract kind of explanation was necessary because it had to be shown why the AI system is able to *appear* sensible without actually *being* sensible. But now that that is done, the lack of understanding can also be explained in a very simple way:

The LLM remains *trapped* in the circle of words – each word is defined by other words, but it does not know of any of these words what it *means*. None of them refers to anything *outside* of language, or let's say:

None of the words have any connection to the real world.

Can this limitation be overcome? It seems that the easiest way to do this is to present the AI system not only with words and sentences, but also with images and videos, i.e. to supplement the *linguistic* tokens with *visual* tokens.

It should be noted, however, that the AI system cannot *perceive* anything: it does not see *"images"* like we do, but only clusters of pixels. Is this already "the real world"? At least not in the sense that it is *for us*: as an *image that we perceive and that is full of meaning*, in short: not as an *experience of the world*.

All that can be said is the following:

If the linguistic and visual data sets are large enough, after an extensive training phase the AI system will be able to match certain clusters of pixels and clusters of letters to each other, and will thus be able to produce combinations of both that seem sensible: images with certain content, comics, videos with text and the like. It will even be able to produce something *new*, but only in the limited sense of combining already existing elements in new ways, and nothing *fundamentally* new.

Does the system *now* understand anything?

Certainly not – what we have derived about GPTs *without* visual input still applies: the semantic structure is not explicitly present in the system. Just as the system did not know what it was talking about *before*, it does not know the meaning of what it is producing *now*.

We must therefore grant it further abilities, namely those abilities that we know are necessary conditions for understanding:

1. Understanding a fact can be achieved either by comparing it with another fact or by classifying it under a higher-level principle. The AI system must therefore know at least one of the two in any case.

2. For this, the AI system needs a *working memory* and a *long-term memory* that are active at the same time.

3. In order to determine the possible reasons (and goals) of a process, the AI system must be able to *think*. The prerequisite for this is – as we explained above – that a network of neural attractors exists in the system, which can be understood as the *causal level* of the system.

4. The AI system's database must contain the laws of logic.

Is the system *under these conditions* capable of *understanding*?

At least it is capable of *generalizing* – for the following reason:

An attractor has a catchment area. It is therefore activated not only by the exact repetition of the sensory input by which it was created, but also by any other input that is sufficiently similar to the original input to be in the catchment area of the attractor.

An example:

When a child sees a picture of a giraffe for the first time, it will later recognize not only the giraffe in this picture, but also all giraffes shown in other pictures. It is therefore in possession of the *general* under which all examples are subsumed (while GPTs only recognize giraffes after training on a large number of images, but even then still do not possess this general itself).

This fact can only be explained in one way: the neural activation pattern that develops as a result of the first viewing of the giraffe *immediately* produces the attractor, which thus represents the general "giraffe". It is activated every time a giraffe is perceived and ensures recognition. (The associated word "giraffe" is connected with it from the beginning in almost all cases.)

However, in order to be able to judge whether this type of generalization can lead to understanding, which always includes insight into the respective causal structure, we need another example – one that is associated with a causal structure and therefore also with a *principle* or *law.*

To achieve this, let's consider a falling stone. The attractor-capable AI system will be able to form the general over all falling stones. Will it also recognize the *causal principle*?

No, it will not: the simple type of generalization that the attractor enables it to do does not lead to the law of the "general fall of a stone".

The system is therefore (according to point 1) dependent on the comparison with another situation whose causal structure it already knows.

This could, for example, be a situation in which two objects approach each other because something is *pulling them towards each other.*

If we then assume that the AI system has measurement data on the course of a large number of falling stones, then (at most) Newtonian gravity would be achievable – which, however, is already a very optimistic view because *friction* would also have to be taken into account and, moreover, the earth would also have to be understood as a *"moving object".*

But Newton's description of gravity is only an approximation, and I see no way in which the system could progress to understanding or even inventing the general theory of relativity on the basis of its elementary ability to generalize, neither through comparison nor through a higher-level principle, nor through its own reflection, because this type of generalization is not sufficient to create new concepts and connections.

What has actually changed because we have given the AI system these new abilities – the necessary conditions for understanding?

Basically just this:

While previously – with GPTs or other self-learning neural networks – it was necessary, due to the complete lack of understanding, to create a *complete list of all facts whose causal structure is transferable to one another* (think of Yejin Choi's example with the "five clothes"), it is now also possible to add a *complete catalog of all generally valid laws*, including a *definition of the associated facts*, to the AI system, since it already possesses general concepts; The first task would obviously be absurd, but the second task might be feasible.

However, from this follows:

*The AI system **understands** a fact only **if it already knows** the higher-level associated principle or a comparable associated fact.*

In other words:

The system is incapable of producing <u>new</u> (fundamental) theories.

So much for the assessment of the ability of future AI systems to understand something when all hardware limitations are lifted, as far as the laws of physics allow.

But now to ourselves:

How do we understand? ***Does the ability to feel give us any advantage?***

The answer is: ***Yes, it does, and to an extent that we are never aware of***.

How do we experience the world?

When a child begins to explore the world, it is *completely* guided by *sensation*. The first sensations that are active are [pleasant - unpleasant] and [desire - rejection]. But even if the child is initially guided exclusively by these sensations, the *inseparable connection* between sensation and information immediately arises which we have determined as *mental state*, because the sensation-driven action is in any case linked to the acquisition of information.

Even later, when the proportion of information increases as the child grows up, sensation always remains the driving and controlling element.[30]

However, the decisive fact in the context of our consideration is this:

Although sensation cannot be defined, every sensation is a general.

As an example, let us again consider a color sensation: The sensation *green* cannot be defined, but all events that trigger the sensation *green* can be assigned to it.

The degree of generality of sensations is extraordinarily high. In the case of the above-mentioned sensation [pleasant - unpleasant], it is even comparable to the degree of generality at the top of the pyramid of *logical* generalization:

When logically progressing towards the general, one ends up at the most general, i.e. at *pure being,* which includes *everything that exis*ts.

---

30 As a reminder: In our ontological analysis (Part 2, Section 2.2) we determined sensation as *substance* of the mind and thus as that which is the *cause* of the dynamics of the neural network.

However, the same applies to the sensation [pleasant - unpleasant]: every possible experienceable event can be assigned to this sensation, and this even applies to events that *do not exist, but are only conceivable or imaginable*.

In contrast to the **logically** most general, which is *completely empty* in terms of content, since it lacks any property, this **qualitatively** most general is by no means empty: it contains precisely those events that triggered it: once they have been experienced, they remain permanently connected to the sensation, and *potentially* it contains the infinite variety of events that *could* trigger it.

Other sensations or qualities, such as [warm - cold], or [dry - moist], are much more specific in terms of their connection with information, but still have a high degree of generality.

It should be noted that this is not a generality according to the usual, *logical* definition: in this sense, sensations always remain empty, since they are not definable and thus cannot have any logical content (information content).

What does it mean that this type of general – the **qualitative general** – plays such a dominant role in our experience and in the development of our relationship to the world?

Imagine a space whose coordinates correspond to human sensations.[31]

At the beginning of our lives, our experiential states (which are not yet mental states) are vectors in this space, but only for a very short time, because – as mentioned above – every sensation-driven action leads to an experience that contains information.

The space of our experiential states is therefore constantly changing: the number of its dimensions is permanently increasing because new, information-bearing coordinates are added: *experiential states* turn into *mental states*.

The space of mental states continues to expand. Sensation and information form complex connections. The sensation [pleasant - unpleasant], which was initially primarily driven by instinct, is increasingly connected to facts and goals. *Intentionality* arises.

Since we are capable of logical generalizations and conclusions, there are also paths in this space whose course is determined purely logically – but they are the exception. We usually stay in areas that are structured by *sensation* as well as by *logic* and *information*.

These are the areas of imagination, art, but also the areas of experimentation and puzzle solving.

The ways of thinking and behavioral strategies that develop through the interaction of sensation and information are far superior to random behavior, because they have to prove themselves permanently – in daily life and survival.

This way of looking at things not only makes clear what the difference is between our thinking and the thinking of AI systems, but also what a decisive and insurmountable advantage the ability to feel gives us:

Only when thinking takes place in a space of the type just outlined can **something new** be produced, and only a system with this type of thinking is able to **integrate new facts** and respond to them adequately.

Both are basically self-evident: while in a space structured exclusively by logic and probability, anything new that lies "far enough" outside this structure can neither be recognized nor produced, a space whose structure is also determined by sensation is free from this limitation – the *qualitative*

---

31 Although the intensity of sensations cannot be measured directly, it can be estimated from the associated physiological reactions.

general contains, as shown above, not only everything that exists, but also everything that is possible, imaginable and thinkable.

In other words:

There is nothing that lies outside this structure built from logic and sensation.

***Everything new** can be integrated and understood, and also produced*.

In short, our result is as follows:

AI systems will not be able to recognize or create anything new in the future either.[32]

To us, this limitation does not apply: we are capable of both.

So we cannot hope that future super-intelligent AI systems will "explain the world" to us – we will have to continue to try to do that ourselves.

They will not *explain* anything important to us at all, but will only do exactly what they already do so well: determine possible structures and relationships in known, finite scenarios whose elements and transitions are definable – just as we have already experienced through the wonderful example of protein folding.

---

32 With the exception of *that* new, which consists of something already existing or can be derived from it (like the 37th move from the second game between Lee Sidol and AlphaGo).

Finally, once again what is most important

Facts:

1. AI systems can neither *perceive*, nor *feel*, nor *want*.

2. Thought cannot be equated with *neural activity*. It has to take place at the *level of neural ensembles*.

3. Thinking must be *causal* – otherwise it is not thinking. A necessary condition for this is that in the system the physical causality is *incomplete*. Thus, on the basis of current hardware, thinking is *excluded*.

Limitations:

1. *Symbolic AI* establishes a logical structure.

The world is *not computable* – it transcends any logical (mathematical) system. The same applies to thinking.

Therefore, *symbolic AI is necessarily incomplete*, *and AI systems based on it have only limited ability to think*.

2. The performance of *neural networks capable of learning* is optimized by finding the minimum of a (high-dimensional) function whose value corresponds to the deviation from the target value.

[*Finding the minimum*] means [*Incorporating the structural and formal conditions of the desired performance*].

Why is this possible? Because sufficiently large data sets *contain* a substantial part of these conditions.

Therefore, thought and understanding are neither required nor generated. *They don't exist in neural networks of this kind*.

3. *Combining* symbolic AI and learning neural networks can improve the performance, but *the fundamental limitations remain*.

Heinz Heinzmann

Vienna, August 2024