Why We have Free Will and Consciousness and AI Systems do Not

Contents

1. The Proof 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 assumed, for example, that we expressed a thought B ***because it follows from another thought A***, then that would be a self-deception. It would mean postulating a causality at the level of mental processes, in other words: a causality from "above". The idea that thinking itself leads to correct results obviously *presupposes* its causal effect. How else would it be possible to correct an error *intellectually*? – If my *thinking itself* were not causal – would *physics* correct itself?

You have to decide: Causality lies ***either*** in thought ***or*** in physics – both at the same time are not possible: B would then be "causally overdetermined".

From the reductionist point of view, 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.

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:

Proposition:

There is a fundamental difference between a really existing system and its representation: the really existing system is active, but the representation is not active.

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. [1]

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 not 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.

1 It would be more than absurd to assume that a blade of grass ***calculates*** where it should move – it simply follows the wind that touches it.

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 causal level? – Or, to put it another way: Which level is dominant?

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.

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.[2] 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.[3]

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 first note on page 21)

2 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.

3 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[4] 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.

4 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,[5] 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.

5 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."

Proposition:

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.

Note:

Of course this does not mean that volitional decisions cannot be analyzed with respect to their neural, chemical, physical, genetic, social, psychological 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 not 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 glass vessel.)

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.

For the following proof that *we ourselves* have sensations and consciousness, while *AI systems* remain insentient and unconscious, the results derived in the proof of free will are presupposed.

2. Why We are Sentient and AI Systems are Not

2.1. Ontological Prerequisites

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

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

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

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. But **substance is always presupposed:** We know that objects are **activated** by ***mass*** or by ***charge***, but we do not know what mass and charge "**are**".

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. In the Section on Free Will, we have proven that the mental level is the ***causal level***. So we are no longer in the physical realm, and this means that 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: *information* and *sensation*.

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

Sensation must be understood here in the broadest possible sense: It 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:

Sensation **is what drives the dynamics of the mind**.

2.2. Execution of the Proof

Now we are sufficiently prepared to explain why we have sensations and AI systems do not.

First we need the following

Definition:

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

(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 (the* ***substance****) 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*** takes the following form:

The dynamics of a simulation – contrary to the original – 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 therefore *not* activated by 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

<u>Proposition</u>:

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***.[6])

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.

6 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?

It applies:

Substance and accident always form an ***inseparable unity***.

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

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.

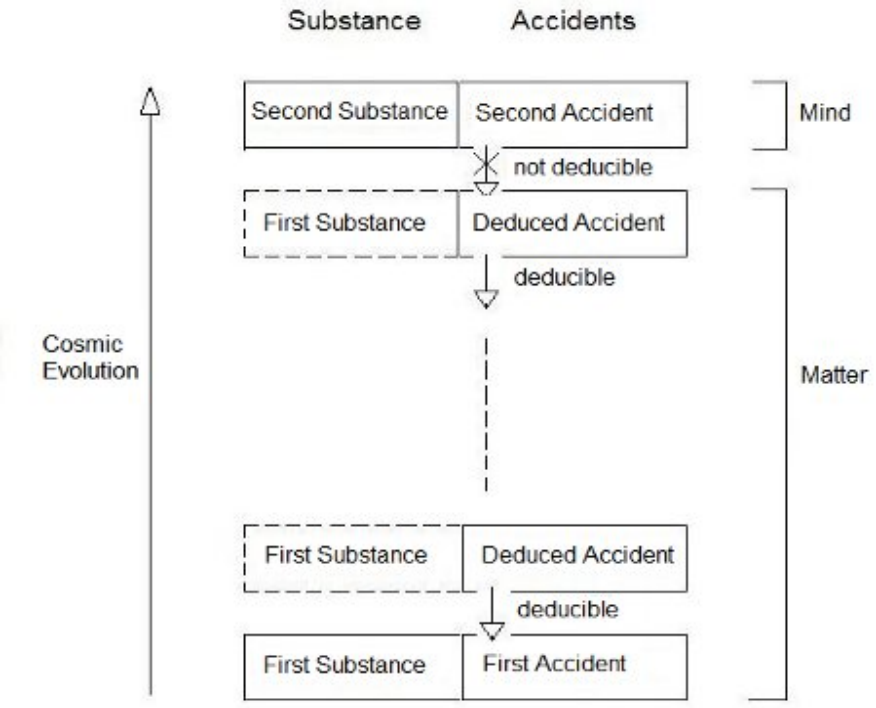
Due to the ***inseparability*** of first substance and first accident, the following applies:

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 <u>different</u> from the first accident, then the associated substance must also be <u>different</u> 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.

In ourselves, this condition is fulfilled: the substance is transformed – we have sensations.

But the dynamics of a simulation is based on *supplied* activity. Thus, the accidents are *not* activated by 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.

The question is: How far does our proof that AI systems cannot possess consciousness extend?

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 and definable – 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 definable.

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 <u>more</u> 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:

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:

Proposition:

It is not possible to construct an AI system that experiences sensations and has consciousness. Neither in a simulation nor in a replica 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.

2.3. Sensation and Artificial Intelligence

The usual question is:

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

We asked ourselves the question instead:

"Why does the indefinable, which exists everywhere in reality, 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 – as we have shown – it can be found *in everything that exists* and is therefore self-evident [7] – 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*.

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***.

I have defined the term "sensation" differently from its usual meaning. I would like to explain in a little more detail why this was necessary and what follows from it:

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, 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 for illustration: *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 as a whole – as an "image" – and on which we see, hear, feel, smell and taste.

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*.

7 See page 12.

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 made up of sensations.

So to put it very clearly:

AI systems do not see, do not hear, do not feel, do not smell, do not taste.[8]

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. However, pixels that only transmit ***information*** about brightness and color cannot be combined to form an image, unlike the *same* pixels when their content is ***perceived*** as brightness and color: it is immediately clear that they can then be added together to form an image.

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*.[9]

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 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, 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.

---

8 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.

9 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.

In summary:

1. ***AI systems cannot perceive 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 experience anything***.
   They have no feelings.
3. ***AI systems cannot want anything***.
   They lack intentionality and motivation.

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.

Note:

Our proof also refutes the so-called "simulation hypothesis": if our reality were a simulation, then *we ourselves* would have no sensations and no consciousness.

Note:

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 above) 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:* 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 (desire, longing, suffering, empathy, etc.) 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:

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[10] – just as we assume for human neural networks – 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.

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

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.

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.

Here is an example which demonstrates both aspects of the attractor hypothesis: recognition after only one encounter and ability to generalize:

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.

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:

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 by 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.

Heinz Heinzmann

Vienna 2023

There is also a longer version of this paper in which the consequences for AI are analyzed.