

### 3. What follows for AI? <sup>1</sup>

#### Contents

3.1. Introduction.....	1
3.2. What is "sensation".....	1
3.3. What can be ruled out with certainty .....	3
3.4. Which limitations are likely .....	4
3.5. Overview, comparison, final assessment.....	17
Finally, once again what is most important.....	25

#### 3.1. Introduction

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 its existence is self-evident<sup>2</sup> – 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

---

1 For the following, the contents of [Parts 1 and 2](#) of this paper are presupposed.

2 See Part 2, [page 16](#).

"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*.

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.***<sup>3</sup>

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 *qualie*, i.e. consists not only of *information* but also of *sensation*.<sup>4</sup>

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

---

3 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.

4 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.

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, 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."<sup>5</sup>

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.

---

5 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.

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.

In summary:

***AI systems are not a new, super-intelligent, dominant species. They are not living beings, but automatons.***<sup>6</sup>

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:

---

<sup>6</sup> 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*.

"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."<sup>7</sup>

(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.

---

<sup>7</sup> 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 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."]

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.

"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 (in Part 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 mass guides the objects of a gravitational system and connects them with each other – like e.g. Earth and Moon – so sensation guides the objects of a mental system and connects them with each other – like e.g. child and ball – and it does so, like mass, by itself.***

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.<sup>8</sup> 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.

---

<sup>8</sup> 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.

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?*

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 not 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.<sup>9</sup>

---

9 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.<sup>10</sup>

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,<sup>11</sup> 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.

---

<sup>10</sup> <https://arxiv.org/abs/2211.00241>

<sup>11</sup> 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:<sup>12</sup>

"...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.

---

<sup>12</sup> 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.<sup>13</sup>

---

13 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.<sup>14</sup> 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!"

---

<sup>14</sup> 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.<sup>15</sup> 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.

---

<sup>15</sup> 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*.<sup>16</sup>

This means:

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.

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.<sup>17</sup>

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.

---

16 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*.

17 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 Minsky's [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.<sup>18</sup>

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 really 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 causal level of the neural network, and that in turn is only possible if physical causality is incomplete.***

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?

---

<sup>18</sup> 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.<sup>19</sup>

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

---

<sup>19</sup> 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, [page 17](#) 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 new (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.<sup>20</sup>

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 exists*.

---

<sup>20</sup> 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.<sup>21</sup>

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*

---

21 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.**<sup>22</sup>

**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.

---

<sup>22</sup> 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.

***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