Filozofski vestnik
ISSN 0353-4510 | Volume 45 | Number 2 | 2024 | 301–28
 | DOI: 10.3986/FV.45.2.14

Aleš Bunta*

Artificial Intelligence as a Metaphysical Event: The Problem of Understanding[1]

Keywords
artificial intelligence, understanding, Nietzsche, Hinton, embodied errors, micro-evolution

Abstract
The paper focuses on the questions of whether, to what extent, and in what ways the implications of the rapid development of artificial intelligence are changing the nature of one of the fundamental philosophical questions, "What does it (even) mean to understand?" It draws on two sources in particular: Hinton's explanation of the technological development and functioning of deep neural networks and Nietzsche's deconstruction of human understanding based on his key concept of "embodied errors." In doing so, it reveals a series of unexpected parallels, relating in particular to the notion of micro-evolution and the function of error in the processes underlying "thinking" and "intelligence." The paper therefore draws certain parallels and demarcation lines between human understanding and the "learning" procedures of digital neural networks. At the same time, it addresses the question of what it means for the interpretation of human understanding that, for the first time in history, understanding is faced with a real, existing antithesis, represented by intelligent systems which, although they do not understand, are capable of performing the tasks of understanding, and capable of replacing understanding.

[1] This article is a result of the research programme P6-0014 "Conditions and Problems of Contemporary Philosophy," and the research project J6-4623 "Conceptualizing the End: Its Temporality, Dialectics, and Affective Dimension," which are funded by the Slovenian Research and Innovation Agency.

* ZRC SAZU, Institute of Philosophy, Ljubljana, Slovenia
ales.bunta@zrc-sazu.si | https://orcid.org/0000-0002-2901-2692

Umetna inteligenca kot metafizični dogodek: problem razumevanja

Ključne besede
umetna inteligenca, razumevanje, Nietzsche, Hinton, utelešene zmote, mikroevolucija

Povzetek
Prispevek se osredotoča na vprašanja, ali, do katere mere, in na kakšne načine, implikacije pospešenega razvoja umetne inteligence spreminjajo naravo enega temeljnih filozofskih vprašanj, »kaj (sploh) pomeni razumeti?« Opira se zlasti na dva vira: na Hintonovo pojasnjevanje tehnološkega razvoja in delovanja globokih nevronskih mrež in Nietzschejevo dekonstrukcijo človeškega razumevanja, ki temelji na njegovem ključnem konceptu »utelešenih zmot«. Pri tem, odkrije serijo nenadejanih vzporednic, ki se nanašajo zlasti na pojem mikroevolucije in na funkcijo zmote v delovanju procesov, ki tvorijo podlago »mišljenja« in »inteligence«. Prispevek torej potegne določene paralele in demarkacijske linije med človeškim razumevanjem in postopki »učenja« digitalnih nevronskih mrež. Obenem pa se ukvarja z vprašanjem, kaj za razlago človeškega razumevanja predstavlja dejstvo, da je slednje, prvič v zgodovini, soočeno z realno obstoječo antitezo, ki jo predstavljajo inteligentni sistemi, ki so, čeravno ne razumejo, sposobni opravljati naloge razumevanja in sposobni razumevanje nadomestiti.

∞

I would like to draw attention to a parallel that can be drawn between the form of genuine philosophical questioning and the form of one of the central questions posed by the emergence of artificial intelligence. This parallel carries a message: it implies that the challenge addressed to our civilization by the rapid development of the technologies that drive artificial intelligence *can be articulated* in the form of a philosophical question. This is not to be taken for granted: technological developments largely escape the grasp of philosophical concepts, at least until their impacts reach us in full force.

It is important to emphasize that this parallel is a formal one—it refers to the characteristics of the question, not to its content—but at the same time, almost in the same breath, to also add that the parallel nevertheless cannot be reduced to a simple structural overlap or described as a search for related patterns in a given database, as AI developers would put it.

This is evidenced by the fact that the central role in this comparison belongs to a certain concept of an event. Heidegger's conception of genuine philosophical questioning, on which I rely here, can also be described as a theory of an intra-philosophical event. For according to Heidegger, the essential distinctive feature of the genuinely philosophical question is precisely that the philosophical question is capable of provoking within itself what he calls a *Geschehnis* (a "happening").

A *Geschehnis* lies, for Heidegger, strictly in the domain of a single, content-specific question, in which the fate of metaphysics is concentrated from its pre-Socratic beginnings to its consummation in Nietzsche's philosophy. This "first," "deepest," "most essential," "most fundamental" of all philosophical questions, according to Heidegger, came to the surface in its clearest form in Leibniz's famous question, "Why are there beings at all instead of nothing?,"[2] which Heidegger, in an essential shift, abbreviated to the question "Why?"

Since a *Geschehnis* in Heidegger is essentially related to the content of the question, it is in its own way unexpected that he himself has done the work for us and introduced the concept in an extrapolated, formally refined form, without any direct reference to the content. However, in a sense, this is also understandable, because a *Geschehnis* is not, by any of its characteristics, an *answer* to the question "Why?" One could almost say the opposite, for Heidegger describes a *Geschehnis* as a kind of recoil, a "*Rückstoß*" of the question *from its content back towards itself*. Let us pay attention to both the evental dimension as well as to the performative features of Heidegger's introduction:

> But *if* this question is posed, and provided that it is actually carried out, then this question necessarily recoils back from what is asked and what is interrogated, back upon itself. Therefore this questioning in itself is not some arbitrary process but rather a distinctive occurrence that we call a *happening*.[3]

As the emphasis on eventfulness is more pronounced, I add the original:

---

2 Martin Heidegger, *Introduction to Metaphysics*, trans. Gregory Fried and Richard Polt (New Heaven: Yale University Press, 2000), 5.

3 Heidegger, 6.

> Aber *wenn* diese Frage gestellt wird, dann geschieht in diesem Fragen, falls es wirklich vollzogen wird, notwendig ein Rückstoß aus dem, was gefragt und befragt wird, auf das Fragen selbst. Dieses Fragen ist deshalb in sich kein beliebiger Vorgang, sondern ein ausgezeichnetes Vorkommnis, das wir ein *Geschehnis* nennen.[4]

A *Geschehnis* thus unfolds as a kind of counter-impact that the question (regarding the meaning of Being) provokes only if it is truly "posed," truly "carried out," and which causes the question to recoil back—from what is questioned within it (the meaning of Being)—into itself and call itself into question.

Since it at best leads to the questioning itself becoming part of what is questionable, a *Geschehnis*, at least at first glance, contributes nothing to the solution of the question and certainly does not provide an answer to it. In reality, however, something even more fundamental happens with its occurrence: although a *Geschehnis* does not provide an answer, it enables the question (about the meaning of Being)—by calling it into question—*to be genuinely posed to begin with*.

> Why the Why? What is the ground of this why-question itself, a question that presumes to establish the ground of beings as a whole? Is this Why, too, just asking about the ground as a foreground, so that it is still always a *being* that is sought as what does the grounding?[5]

Clearly, "Why the Why?" is not simply a consideration of the validity, the legitimacy of the "Why?" question itself. The questions "Why?" and "Why the Why?" work together, one with the other, one against the other. And in fact, it is only somewhere in between them—provided that a *Geschehnis* has taken place and the original question is placed in the vicinity of nothingness—that the dimension of meaning, *the meaning of Being*, is revealed; now no longer as a question, but as itself.

The above could hardly be further removed from the topics of deep neural networks, the backpropagation algorithm, the promise and risks of the unprecedented capabilities of intelligent systems; further removed from the profound

---

4 Martin Heidegger, *Einführung in die Metaphysik* (Tübingen: Max Niemeyer Verlag, 1953), 4.

5 Heidegger, *Introduction to Metaphysics*, 4.

societal changes these systems carry out, the dangers of their misuse or potential escape from human control, or the theory of singularity—all of which are a part of what we commonly refer to as the dispositive of artificial intelligence.

And indeed, my initial thoughts had no better ground than a risky intuition that could not be dispelled. Namely, that in philosophical reflection on the question of artificial intelligence, a similar counter-impact is triggered, one that would strike at the very essence of philosophy, but in the area of a different, no less fundamental philosophical question than "Why are there beings at all instead of nothing?" This question is: "What does it (even) mean to understand?"

One of the lessons of Heidegger's treatise is that fundamental questions are raised when they are confronted with a void in themselves; when they are called into question. And I think that, in the broadest sense, this also applies to the contact of the question of the meaning of understanding with artificial intelligence.

The topological relocation of a *Geschehnis* to the realm of understanding is obviously also risky at the conceptual level. Heidegger's notion—therein lies to a large extent his wager—operates in the pure immanence of questioning. The game of question marks is ultimately the only one that stands in an inner relationship to the sense of Being that is no longer merely a question mark. By contrast, the contending *Geschehnis* that—at the level of the question "What does it (even) mean to understand?"—is supposedly triggered by the attempt to understand artificial intelligence, is obviously open to the outside, it is triggered from the outside, by some factual development, which makes it seem that the whole problem might be more appropriately addressed by Badiou's pairing of a scientific event and event fidelity (philosophical interpretation).

To start with Badiou: I simply do not believe that the AI dispositive can be captured by the notion of truth-event; I also doubt that Badiou would have wanted such a thing. It seems to me more reasonable, and fascinating in its own way, to suppose that in this unfolding the event is absent, or rather diffused. As for the problem of immanence, my answer is twofold. Throughout this discussion I will show that—even though a *Geschehnis* unfolds in the field of understanding in relation to the outside—a key role is nevertheless played by an inner, introverted negative capacity of the very question "What does it (even) mean to understand?," which makes it akin to the question "Why?" On the other hand, I will

show towards the end of this discussion that also Heidegger's wager on the pure immanence of questioning, especially when it is deployed on the historical level, nevertheless needs the intrusion of a necessary disturber, a minimal external trigger, which in Heidegger's treatise is represented by the signifier Nietzsche—the one who actually thought, or in Heidegger's vocabulary, *essentially experienced* that Being has become merely a deflated word.

The Brain and the Microevolution

Do intelligent systems such as ChatGPT *actually understand*? Some scientists, and especially philosophers of science, who tend to be more rigorous in this respect, would say that the question itself displays a superficial understanding of the problem and is evidence that one has fallen for the hype. They may be right, but the question still seems legitimate. Especially given the fact that the recent development of artificial intelligence has revolved in every conceivable way around *the key concept of learning*, a concept which—of all those concepts that would in normal circumstances be said to denote forms of thought—is the most congenial to understanding, or at least seems to be intrinsically linked to understanding in a form of co-dependency.

That intelligent systems *learn* in a completely untransmitted sense, as well as that they are capable of learning on their own—these are no longer dilemmas, but facts. For example, in the course of completing a task, if this helps to solve it, they can easily learn Chinese—in normal circumstances, we would say they learn to *understand* Chinese—without being told to or taught Chinese by anyone. However, learning is not only a capability that intelligent systems possess, but also the name of the development process (so-called machine learning) that, in the strictest scientific and technological sense, produces the very effect that is referred to as the "*intelligence*" of a system. Even in this case, it is not just a technical term, but the word "learning" actually provides a surprisingly accurate description of the process. Therefore, the question could also be posed in the following way: Is it possible that an intelligent system that is *essentially determined by learning* does not understand at all? We must at least concede, without doubt, that—except in the famous case of learning by heart, but even then, not really—a minimum degree of understanding is spontaneously perceived as an internal condition of learning, and that the relationship between the concepts cannot be intuitively explained in any other way.

While there is no perfect consensus among scientists, their position on the dilemma of whether intelligent systems understand is, in the vast majority of cases, reserved: maybe one day, but not yet.

Summarizing their arguments roughly, this reticence certainly appears justified. What, at the deepest level, forms the basis of the learning and "speaking" of systems such as ChatGPT is in fact a complex statistical system that can—on the basis of the complex but, in terms of epistemological status, nevertheless purely and exclusively statistical processing of enormous amounts of data—predict, anticipate, every next word in a sentence, without this meaning that it really understands its content. In other words, although ChatGPT "speaks," the "thinking" that drives this speech is *entirely heteronomous to the meaning* that is established (for us and exclusively for us) within what is uttered. Rather than standing in any focused intimate relationship with the meaning of the sentence, the intelligent system remains at all times dispersed in a relation to the totality of everything ever posted on the Internet, from which it statistically induces a prediction of the next word through a series of contextual parameter constraints. In this respect, intelligent systems do not "understand" any more than a pocket calculator does: they calculate and serve up understanding, but they do not understand. In fact, considering that a pocket calculator, so to speak, stands its ground, it *calculates*—we could even speculate that an intelligent system understands *less*.

But is there not a paradox in saying that something that speaks only speaks *apparently*—especially if we consider that intelligent systems do not merely perform a morphological imitation of words, which is characteristic of some animal species, but carry out a process of *anticipation* which, even if it is statistical in design, in a sense belongs to the category of thinking?

This brings us closer to the argument of those scientists—including two key figures in the recent development of AI, Geoffrey Hinton (also called the "godfather of AI") and Ilya Sutskever (chief scientist in the development of ChatGPT)—who, contrary to most, argue that the first traces of understanding can already be seen in the workings of digital deep neural networks at this very moment. Namely, their argument, which is much more multifaceted than can be summarized here, can be described in the most basic terms as reaching a conclusion on the basis of the effect: Hinton and Sutskever argue that intelligent systems

are capable of solving certain complex problems which, in their view, simply cannot be solved except on the basis of their contextual understanding.[6] This conclusion is then further reinforced by the fact that at the level of the so-called hidden layers of neural networks, there are certain unforeseen qualitative leaps taking place (the so-called black-box effect) that we cannot really explain yet. This is also one of the main reasons why, according to some experts, it is already possible to speak of the autonomization of intelligent systems: their functioning in some cases gives the impression of having escaped from the matrix of their statistical design.

It is not up to us, of course, to intervene in this debate; much more relevant is another, much simpler observation: all those scientists who, in one way or another, answer the question of whether intelligent systems understand, must do so on the basis of some *understanding of understanding*; each of them must, if not otherwise spontaneously, respond to the philosophical question: *What does it* (even) *mean to understand*?

The question of the meaning of understanding, however, is far from being included in the field of artificial intelligence only through interesting, but scientifically probably nonetheless trivial questions. First of all, we should note the following: the question "What does it mean to understand?" is the philosophical parallel to the *idea* that set technological development and the effects we are witnessing today in motion to begin with.

The original aim of the teams of scientists who set out to develop and research the so-called digital neural networks that form the heart of the most advanced intelligent systems currently in use was precisely this: *to build a system of understanding the workings of the biological, human brain by means of a digital reconstruction of it*, which, due to its mathematical design, can be understood better and more efficiently than (our own) biological brain, the principles of which in many respects still remain unknown. Hinton, who at the beginning of his rather

6 Sutskever, on some occasions, puts forward the argument even more directly: "To predict the next token means that you understand the underlying reality that led to the creation of this token. It is not statistics." Ilya Sutskever, "Why Next-Token Prediction is Enough for AGI," YouTube video, uploaded by Dwarkesh Patel, December 13, 2023, 00:58, https://youtu.be/YEUclZdj_Sc.

turbulent university years also studied philosophy—a not inconsiderable fact—thus said in a recent public lecture held at Cambridge University:

> I was very interested in philosophy of mind, but actually it was then when I was doing philosophy when I was about nineteen that I formulated this view that subjective experience is just shorthand for "I'm going to talk about how the world would have to be to explain what's going on in my head as normal perception," but they weren't too interested in that, so I actually have a grudge against philosophy. [. . .] So, then I decided: you'll never understand how the brain works unless you build one. This is Feynman's view, Feynman wrote this somewhere.[7]

Clearly, there is a huge gap between the original idea and the discovery, invention, innovation, mathematical solution that ultimately makes the idea work—even more so in technology than elsewhere—which often leads to a deviation in a completely different direction than the one planned. This was, as Hinton himself often emphasizes, also the case with deep neural networks. There is a peculiar irony in the fact that the very mathematical invention whose implementation had been pioneered by Hinton, Yann LeCun, Yoshua Bengio, and others in the mid-eighties—the so-called *backpropagation algorithm*—which subsequently, as computers became more efficient, allowed intelligent systems to become truly similar to us, through their ability to learn and the effect of language, on the surface, has also caused intelligent systems to diverge significantly from the direction of the way the human brain works at deeper levels. Namely, this algorithm, which forms the basis of the training of advanced digital neural networks, operates according to principles that are completely different from those of the biological brain.

The development of artificial neural networks has thus diverged from the original idea. Besides that, we also have to admit that the parallel between the idea of understanding the human brain on the basis of its digital, mathematically "manageable" reconstruction, and the philosophical question "What does it (even) mean to understand?" turns out to be rather naive idealism at the level of neuroscientific practice. Understanding is, of course, only one of the many cog-

7 Geoffrey Hinton, "Two Paths to Intelligence," lecture at University of Cambridge, May 25, 2023, YouTube video, uploaded by CSER Cambridge, June 5, 2023, 1:06:40, 1:07:39, https://youtu.be/rGgGOccMEiY.

nitive phenomena that the development and research of neural networks is supposed to help us understand. Not only that, understanding is undoubtedly one of those cognitive processes that Hinton referred to by the general term "reasoning" and said were in fact "a bad model of biological intelligence," because, developmentally speaking, they represent *late-developed forms* of brain function.

> Reasoning came much, much later, and we are not very good at it—you don't learn it until you are very old. Reasoning is a very bad model for biological intelligence: biological intelligence is about controlling your body and seeing things.[8]

Understanding of understanding is, therefore, at most one of the secondary and distant goals of computational modelling of the brain. Nevertheless, Hinton's statement, repeated in several interviews, i.e. that "reasoning is in fact a bad model of biological intelligence; it developed much later," deserves our full attention: first of all because it evokes a strong philosophical reminiscence.

In fact, it reiterates words very similar to those—I do not know if Hinton knew this—with which Nietzsche, at the intersection of psychology, epistemology, and genealogy, virtually opened the door to a new terrain, which could be called the *microevolution of the human being*. For Nietzsche, too, argued precisely this: that processes and concepts such as cognition, understanding, intelligence, knowledge, even consciousness to some extent, and above all truth, are "too young," too "late-born," too late, of secondary origin, to allow us to pin all our hopes on them.

Nietzsche's thesis is in fact twofold:

a) Processes and fundamental concepts such as cognition, understanding, knowledge, intelligence, and truth actually emerged late in the process of the micro-evolution of human becoming; too late, too derivative of other, more elementary processes, to be counted among the key factors forming the *basis* of human development. They may constitute the culmination of the human being, but they were too late to participate in its *Entstehung*, in the stage of human for-

---

8 Geoffrey Hinton, "The Godfather in Conversation: Why Geoffrey Hinton is Warried About the Future of AI," YouTube video, uploaded by University of Toronto, June 22, 2023, 5:47, https://youtu.be/-9cW4Gcn5WY.

mation. Concretely, when Nietzsche, for example, set forth the sharp formula that "our apparatus for acquiring knowledge is not designed for 'knowledge,'"[9] he meant to say that what we today call the cognitive apparatus was in fact formed in relation to a completely different end, and by means that differ completely from gathering or generating (true) knowledge: the original task of the "apparatus for acquiring knowledge"— one that has stretched over hundreds of millennia of human prehistoric development and can indeed be placed in the structure of human formation— was even something *quite opposite* to acquiring true knowledge, namely, the creation and maintenance of *those errors that have proved micro-evolutionarily necessary for the preservation of the species or for an increase in its power.* The cognitive apparatus, and through it our cognitive, intellectual power, thus evolved through adaptation to *error*, not to true knowledge. The fact that the "human beast," like its animal predecessors, had to learn very quickly to recognize correctly something edible or a danger, confirms rather than contradicts Nietzsche's hypothesis, which puts forward the *conditions for the survival of the species*. Among these necessary, vital, determining errors, according to Nietzsche, belong the notion of the ego, the division of the world into "permanent, unchanging entities," the existence of the will, and finally, indirectly, through the action of language, the notion of being itself. We can see at once that the elementary forces underlying human origin still leave their traces, both in the architecture of language as well as in the structure of thought and, ultimately, in the edifice of consciousness. Although, in the broader context of our discussion, this may be a bolder claim than it would be otherwise, we will say that these necessary errors—Nietzsche calls them "embodied errors"—are instinctively impregnated into our brains.

b) In order to truly understand these late phenomena, which are already "too human" to belong to the origins of humans, and in particular to grasp that inner "tension of the spirit" which undoubtedly belongs to concepts such as truth, knowledge, understanding, and intelligence, we must understand them as well *in relation to the processes underlying their formation*; especially since they have all in their turn *evolved out of their opposites*. The real "tension of spirit" which surrounds and pervades these essential determinants of the category of the human will be sought in vain in some profound spirituality of their origin, nor will

9 Friedrich Nietzsche, *The Will to Power*, trans. Walter Kaufmann (New York: Vintage, 1968), 273.

it be captured by however correct a definition of each concept; it will be discovered in the fact that the opposites from which they evolved remain active just beneath the surface.

Since we will deal with the trinity of understanding, knowledge, and truth in more detail in the next section, for the moment let us just briefly recapitulate Nietzsche's view of consciousness and mind, *Vernunft*, which is particularly interesting because it also includes the meaning of "intelligence." We experience consciousness as an inner state of our existence, the centre of subjectivity. But Nietzsche wrote: "It is essential that one should not make a mistake over the role of 'consciousness': it is our relation with the 'outer world' that evolved it."[10] Consciousness was, originally, only "a network of connections between human and human."[11] Although we experience it as the centre of our subjectivity, it has in fact evolved as an intersubjective process, and underneath the layer of our self-experience, it still remains essentially related to externality.

An even more elementary example of "an emergence from an opposite" is the mind, intelligence: Nietzsche made the seemingly simple but far-reaching claim that at the deepest foundation of every mind, however developed and sophisticated, there is always a "non-mind," an *Unvernunft*. "'Intelligence' [*Intelligenz*] appears as a special form of irrationality [*Unvernunft*], almost as its most malicious caricature."[12] "*From experience.*—The irrationality [*Unvernunft*] of a thing is no argument against its existence, but rather a condition for it."[13]

It seems that intelligence is always established somewhere between itself and an elementary complex simplicity that persists in its foundation at all times and, despite its apparent detachment from it, constitutes it. Interestingly, a similar message is being conveyed on several levels by the discoveries relating to artifi-

10 Nietzsche, 284.

11 Friedrich Nietzsche, *The Joyful Science*, in *The Joyful Science, Idylls from Messina, Unpublished Fragments from the Period of The Joyful Science (Spring 1881–Summer 1882)*, trans. Adrian Del Caro (Stanford: Stanford University Press, 2023), 222.

12 Friedrich Nietzsche, *Nachgelassene Fragmente 1884–1885*, vol. 11 of *Sämtliche Werke: Kritische Studienausgabe*, ed. Giorgio Colli and Mazzino Montinari (Berlin: De Gruyter, 1999), 700.

13 Friedrich Nietzsche, *Human, All Too Human (I): A Book for Free Spirits*, trans. Gary Handwerk (Stanford: Stanford University Press, 1995), 269.

cial intelligence, in particular to the development of neural networks. Although they are extremely complex systems that are being developed on the basis of barely conceivable amounts of ongoing calculation, there is, as the scientists themselves point out, something surprisingly simple at their core:

> When I first learned about this, I was mystified by how something so simple could compute something arbitrarily complicated. [. . .] Although you can prove that you can compute anything in *theory* with an arbitrarily large neural network, the proof doesn't say anything about whether you can do so in *practice*, with a network of reasonable size. In fact, the more I thought about it, the more puzzled I became that neural networks worked so well.[14]

Although we are not quite at the point where we could fully appreciate his profound amazement at the elementary simplicity of how digital neural networks work, this description given by one of the most prominent researchers in the field of AI, Max Tegmark, is certainly intriguing.

But can Nietzsche's lateness of truth and understanding really be related to the developmental lateness of reasoning that Hinton speaks of, and which, according to him, causes reasoning to be a poor model of biological intelligence?

To answer this question, in my reckoning, it is sufficient to place the core of Nietzsche's thesis in direct comparison with Hinton's statement. Nietzsche, as we have seen, makes the following claim: "Our apparatus for acquiring knowledge is not designed for 'knowledge,'" and when Hinton says that "reasoning is a bad model of biological intelligence," is he really saying anything other than that the *brain was not made for reasoning to begin with*? That there is an edifice within it, which originally was not developmentally adapted to the tasks of thinking, which at best form a thin layer on top of it, *which explains precious little*, and which is therefore nearly useless in a research project which has set itself the task of understanding cognitive processes, *including thinking*, at their very core?

[14] Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (London: Penguin, 2018), 74.

Despite their completely different starting points, and undoubtedly also different conclusions and outcomes, Nietzsche and Hinton do meet somewhere. I also believe that this parallel between them cannot be reduced to the "usual" developmental naturalism, through which some authors even associate artificial intelligence with the Aristotelian naturalist outlook. I doubt that Aristotle would have agreed that the processes of thought at their core can only be properly understood by means of *excluding logic from them*—which is what Nietzsche and Hinton explicitly claim. I also doubt that Aristotle, or any other naturalist, would readily accept the idea that the *effect* of correct cognition can be achieved through a process whose *basic building block is error*, while correct cognition, although its effect is ultimately reached, is entirely absent from the process leading to this effect—for this is precisely where I see another, deeper point of contact between Nietzsche's epistemology and the workings of artificial intelligence.

This as yet intangible parallel can be better grasped through three intersections, which I will refer to with the following terms: the biological-microevolutionary element, the negative minimum, and the priority of error—all three intersections are of course connected and intertwined.

The first obvious point of convergence is hence the *wager on biology*. It is clear that the biology Nietzsche is leaning on is not the experimental-technological biology at work in modern neuroscience, which, especially in the field of research into artificial neural networks, slides into mathematics and physics, from laboratory mice to numbers. Biology enters Nietzsche's philosophy through the theory of evolution, through the polemics against Darwin. However, in this very leaning on biology there is another, deeper point of intersection, which I believe is occupied precisely by the notion of microevolution. For the moment, I would like to point out that Hinton in particular appears to insist on the *biological nature of the scientific paradigm* behind deep neural networks,[15] which is in its own way surprising, given the fact that the structure of intelligent systems is purely mathematical. I also think that there is more to this than fidelity to the original, biological brain; on the contrary, I believe that this insistence has almost nothing to do with the biological source.

[15] Geoffrey Hinton, "Will Digital Intelligence Replace Biological Intelligence?," Romanes Lecture at the Sheldonian Theatre, February 19, 2024, YouTube video, uploaded by University of Oxford, February 29, 2024, 4:27, https://youtu.be/N1TEjTeQego.

The second intersection is also quite obvious, but it is covered by a certain appearance of self-evidence. Both Nietzsche and deep neural network research share the view that cognitive phenomena such as understanding can only be adequately explained *through the processes underlying them*. Of course, the composition of this underlying basis varies from one case to another: in Nietzsche, what lies underneath is a historical developmental process driven by the will to power; in computational neuroscience, the underlying basis is the pure immanence of neurons and numbers. It could also be said that Nietzsche and computational neuroscience approach these underlying bases from diametrically opposed directions: neuroscience attempts to explain cognitive phenomena *from within their material under-structures*, which in the case of some cognitive phenomena also poses a problem—according to many philosophers of science, this is particularly true of consciousness, which is itself a "surface effect" of physical processes rather than a physical process, and therefore cannot be explained as such. On the contrary, it is characteristic of Nietzsche's genealogical psychology that it tries to reach the subsurface from within the effects that obscure it. For example, we have seen in the very case of consciousness that its emergence in relation to the exterior has to be accessed through (or rather against) the lived experience of an inner state. Despite all these differences, to which we would have to add completely different means and methods of research, it is nevertheless possible to argue that Nietzsche and computational neuroscience have something in common at the level of approach: Nietzsche and computational neuroscience try to grasp cognitive processes in what I would call the *negative minimum* of the phenomenon.

Someone will say: it is self-evident that cognitive processes must be understood in relation to their underlying framework. But in reality, when it comes to the problem of phenomena such as understanding and cognition, both philosophy and science have for a very long time resorted to a completely different approach; so different that Nietzsche, in his own right, considered that cognition and understanding, and through them the truth, did not really appear before us *as philosophical problems at all* until he managed to turn the perspective on how they ought to be approached, by stumbling upon a different type of question.
The question of idealist philosophy has never been "What does it (even) mean to understand?"; the question has always been "What are the formal conditions of true/correct/objective knowledge?" The question of understanding, for idealist philosophy, is not a question of the underlying basis: idealist philosophy,

through the introduction of a *third aspect*—truth, rightness, objectivity—explains understanding at most as a *means* whose usefulness or uselessness *in achieving the purpose of this third, external term* determines what is and what is not understanding. Knowledge is true or objective knowledge; "false knowledge" is not knowledge at all, but error. Understanding is correct understanding, while "misunderstanding" is a bare privation of understanding—which is de facto contradictory, since the hermeneutists are, I believe, right in this respect: we are *always already caught in an understanding*, beyond the dilemma of right or wrong. Nietzsche was always suspicious of this kind of idealistic use of truth as an external criterion for the categorization of knowledge: he considered it to be an abuse of truth that is particularly detrimental to truth itself. Idealist philosophy does not look to a negative minimum in the basis of understanding, but places all its bets on the *maximum* of true knowledge: it tries to discover its conditions, to set them up as universal, and to define the coordinates of understanding and knowledge through these conditions of reality, of objectivity, of correctness. This is why Nietzsche thought that in idealist philosophy the question, "What does it (even) mean to understand?" (beyond the dilemma of true or false)—despite the appearance that it has always been at the centre of attention—in reality remained unaddressed.

In its own way, it seems even more delicate to claim that science, which is usually assumed to be characterized by the so-called bottom-up approach, has also given up on the material foundation. But the fact is that even in science, taken as a whole, the question of the conditions of knowledge has, until recently, dominated over the question of the processes in the material basis of understanding. And indeed, following Hinton, it can be said that the continuity of this primacy has extended to the terrain of scientific theories and models of artificial intelligence.

Hinton has said that the decisive moment of rupture, which in a sense determined the fate and direction of AI research, and consequently, of course, of its explosive development, resulted from the clash between two models of intelligence that follow two major scientific paradigms—*the linguistic-symbolic* and *the connectionist-biological*—each of which, in its own way, postulates "what is actually inside our heads."

> There are two different models of what intelligence is all about. The first model is all about reasoning. And the way we reason is by using logic—that's special about

> people, and what we should be doing is understanding the kind of logic that we actually use. That also went with the idea that the knowledge we store is *symbolic expressions*. So that I can say a sentence to you, and you will somehow store that and latter you'll be able to use it for inferring other sentences. What is inside your head is something a bit like sentences but cleaned up. And then there's a completely different model of intelligence, which is all about learning the connection strengths in a network of brain cells; and what it is good for is things like perception and motor control. [. . .] That was an entirely different paradigm and it had a different idea of what is inside your head: it is not stored strings of symbols, just the connection strengths. For the symbolic AI view, the crucial question was: What is the form of these symbolic expressions, and how do we do reasoning with them? For the neural net view, the question was quite different: How do we *learn* these connection strengths so you can do all those wonderful things? For the neural net view, learning was always central. For the symbolic AI view, not so: they said, we'll worry about learning later, we must first know how the knowledge is represented, and how we reason with it. So, these are two totally different views: one took its inspiration from logic, one from biology; and for a long time, people from the logic camp thought taking inspiration from biology was silly. This is a bit strange, since von Neumann and Turing both thought that neural nets were the way to attack intelligence, but unfortunately, they both died young.[16]

The symbolic paradigm and the AI models based on it are therefore characterized by the belief that there must be some *minimal symbolic structures* in the brain that allow for at least an approximate correspondence between brain processes and the structure of language, and which, as a consequence of this relative correspondence, also allow for the postulation of an instance *in the brain itself* that guarantees, in the manner of logic, the possibility of the correctness of cognition. The aim of this first model of AI is, as Hinton says elsewhere, "discovering the workings of the logic behind our thinking, which we understand as a distinctive feature of human thought," and transposing this logic onto the functioning of intelligent machines.

By contrast, from the point of view of the biological paradigm, which adheres to the findings of empirical neuroscience, there is nothing in the brain other than synaptic connections. In other words, the biological model of intelligence

---

16 Hinton, "Godfather in Conversation," 4:55, 6:05.

does not postulate any structures in the brain that would guarantee consistency with the forms of thought that take place in the medium of language and that also rely on the structure of language as the criterion of correct cognition, such as logic. For this reason, in this biological model of intelligence, the *concept of learning comes to the fore from the outset*. Since the structure of the brain does not in itself guarantee anything; since there is no *a priori* epistemological criterion, no power of inference, built into it, this can only mean one thing: *all the capacities of the brain must—in one way or another—be learned*.

But what does "learning" even mean in this context? How does the backpropagation algorithm—which, despite being a relatively old invention, remains the key principle of deep learning—actually work? This brings us to the third point of contact, the *primacy of error*.

A very simplified definition might be: backpropagation is a mathematical algorithm that *retroactively calculates the error deviations* in the system's operation, thus allowing the elimination of all those connections in the system's functioning that rank highest in this error coefficient—in short, it allows a gradual serial elimination of all those connections that most strongly steer the system's functioning towards error. In this way, the algorithm, through an almost innumerable number of iterations of the described procedure, gradually leads the system to optimize its performance according to externally defined criteria of correct behaviour/recognition.

> So, the magic is that there's this relatively simple algorithm called backpropagation that takes the error in the output and sends the error backwards through the network and computes through all the connections how you should change them to improve the behaviour. And surprisingly, that actually works.[17]

At first sight, the backpropagation algorithm therefore acts as a means of eliminating error in the service of correct knowledge.

But would it not actually be more correct to say the reverse: neural network learning, according to the principle of the backpropagation algorithm, is a process that continuously relies *on the existence of an error* that is factual and that

[17] Hinton, 12:51.

has to be accounted for, to be excluded from the operation of the system, while correct (re)cognition is added to the process merely *as an effect*—with no correct cognition actually occurring at any point in the process, *including* at its conclusion? To put it in the language of simple ontology: *there is only error* in the process; correct cognition is merely an external effect without any basis of its own—*error brought to almost nothing*.

Let us try to describe the process in our own simplified way, in a little more detail—focusing, of course, on the nuances that are interesting to us.

The first step in training a digital neural network goes something like this: the system is asked a question and responds with an "answer"—I put the word "answer" in quotation marks because, in reality, the system's response has nothing to do with the question, but consists of a purely *arbitrary reaction* that is registered in the system's numerical parameters, which allow its modification. This response therefore contains no knowledge of the question, not even the slightest hint of a correct answer, nothing on which a process of cognition could rely—the irony is that this does not change at all even up to the end of the process.

The important point is that there are several of these absolutely contingent "answers"; they all enter into the numerical parameters of the system, and of course none of them contains even a glimmer of correct knowledge. But even if none of the "answers" contains anything that points to the correct answer to the question, it nevertheless, through its relation to the other "answers," does contain something, namely, a *comparatively measurable degree of its falsity* in relation to what we assume constitutes the correct answer to the question. To put it even more simply, some "answers" are—purely by chance, of course—nevertheless *less wrong* than others, closer to our (external) estimate of the correct answer, and in this triangle between the individual "answers" and what we have determined to be the correct answer it is actually possible to calculate something, namely, a kind of coefficient of error, which allows us, on the basis of this coefficient, to adapt the parameters of the system to those answers which happen to be the least wrong.

In this way, gradually, through an almost infinite series of iterations, we adjust the system, optimize it, until this optimized system, from which we systematically extract the maximum deviations in the direction of error, at some point—

again purely by chance, except that this chance now has slightly narrowed coordinates—does not give a response that *overlaps with what we see as the correct answer to the question*, and then we adjust all the parameters in the system in such a way that this phenomenon of correspondence is repeated as many times as possible. Obviously, this last "answer," despite its overlap with our assessment of "correctness," is no closer to the characteristics of correct (re)cognition than the original, purely arbitrary response; nevertheless, the system, *taken as a whole*, from which we have excluded all the connections that led it into error, begins to behave correctly: it begins to give correct answers of its own accord to many other questions which are not even related to the original one, to behave as if it understood, in short, to produce the effects of correct cognition. I think we can repeat: from beginning to end, the system's agent is error; the "correct (re)cognition" occurs as an *effect* without having taken place.

From here let us return to the comparison with Nietzsche. Of course, it is clear that even if we explain it in this way, the learning of intelligent systems, which erases the parameters of error, stands in diametric opposition to Nietzsche's thesis that thought evolved from the maintenance of fundamental, vital errors, such as the ego and the existent entity; from a kind of coordinated effort to maintain these vital errors, which are necessary for survival. But at the same time, a more fundamental convergence is to be noted: both backpropagation and Nietzsche direct us to the conclusion that *only* "*errors*" *exist* in the processes that form the basis of "thinking."

Let us turn this around in another way. We can observe that, despite the fact that digital neural networks are entirely constructed as mathematical models, there persists in them something—born and emerging from chance—*non-mathematical*. An error is not, after all, a mathematical operation. Of course, mathematics can define it, measure it, and calculate it, but the initial response of the system, which, so to speak, bestows on us the first error, upon which mathematics can then operate, is contingent. And this contingency behind it persists in the process all the time as its central factor, which the mathematical calculation of the error merely selectively directs in the direction of its minimization. This is, after all, at least one of the meanings of the word *learning*: in a sense it denotes the non-mathematical in the midst of the mathematical, *the trace of the biological* that nonetheless cannot be described as a kind of reflection of the actual workings of the biological brain.

In truth, the biological process of learning does seem to be very different. A lot of knowledge is written into us genetically, innately, instinctively. In particular, what Hinton puts at the centre of biological intelligence—motor skills, the functioning of perceptions—are characterized by the fact that they develop spontaneously, without our having to learn them. Then there is the problem of knowledge as such: we humans have to accumulate it, build it; we have to read, contemplate, and deduce. An intelligent system works in the reverse way: first it has all the knowledge of this world, only then can it produce an effect from it that superficially resembles understanding.

So, is "learning" really the best term to describe the emergence of intelligent systems? Should we rather say that this process—which leads the development of intelligence from the zero point of the absolute contingency of the first "answer," through "experience" of the delusions which determine selection, to the effect of correct action—is a kind of *substitute for the evolutionary process*, a kind of micro-evolution accelerated and simplified to the extreme?

Next to Nothing, *Geschehnis*

It is not difficult to see that the question "What does it (even) mean to understand?" diverges from the question of the conditions of true knowledge, and that—especially in the case of Nietzsche's versions of the question—it is conceived through an antagonism with this central epistemological problem. The problem of true knowledge, especially in post-Kantian philosophy, is predicated on the question "What can I know?" Its starting point is therefore the determination of the object of possible knowing, which usually already involves a certain digression from the original ideal (what is knowable is not the "thing-in-itself"; what is objectively knowable is necessarily related to the way in which the subject constitutes phenomenal reality). In any case, what stands at the forefront is the correlation between the object and the edifice of the subject's perceptive and cognitive apparatus.

The question "What does it (even) mean to understand?" on the contrary, focuses on the process of understanding as such—it seeks to discern in it that basic matrix of its operation that is independent of the true/false divide; it seeks to discover what is actually taking place at the moment when we "understand."

Nietzsche himself would say that he is attempting to explain understanding "psychologically."

The second essential difference between the two questions is that the question "What does it (even) mean to understand?" is accompanied by an essential undertone—with a kind of "if anything at all"—which is, however, not to be necessarily seen as an expression of scepticism. Although Nietzsche stressed that scepticism is "healthier" than dogmatism, he nevertheless recognized in it a kind of flip side of idealism: scepticism still proceeds from the idealist conception of "true knowledge," except that it denies it its aspirations.

The nuance of scepticism that accompanies Nietzsche's posing of the question on the meaning of understanding is of a different origin: its source is his, so to speak, preliminary answer to the question of what understanding is. Namely, through his "psychological" consideration, he arrived at two conclusions: first, that the basic matrix of understanding is much simpler than we are willing to admit, and second, that, even at the level of this basic matrix, understanding as such is *indistinguishable from some inherent element of fabrication*. In other words, even if I understand "correctly," I inevitably fabricate, because fabrication is an intrinsic component in the basic matrix of the process of understanding. The blow that Nietzsche dealt to idealism at the level of understanding is therefore conceived not through the denial of the aspiration for truth, through the denial of the possibility of true knowledge, but through a quite affirmative answer to the question "What is understanding?" which is nevertheless such as to call understanding *as such* into question.

> "Inner experience" enters our consciousness only after it has found a language
322 the individual understands—i.e., a translation of a condition into conditions familiar to him—; "to understand" means merely: to be able to express something new in the language of something old and familiar.[18]

> *The origin of our concept of "knowledge."*—I take this explanation from the street; I heard someone from the common people say "he recognized me"—: at which I asked myself: what do the common people want, when they want "knowledge"? Nothing more than this: something strange is supposed to be traced to something

[18] Nietzsche, *Will to Power*, 266.

> known. And we philosophers—have we really meant anything *more* by knowledge?[19]

If we look to the definitions quoted above for the ultimate answer to the questions "What is understanding?" and "What is knowledge?," they are of course disappointing, but that is not their intent. Rather, one could say that Nietzsche is trying to capture in them that negative minimum of understanding and knowledge which, by definition, borders on nothingness. And in fact, the basic message of both definitions—their simplicity is also included in this—is precisely this: at the level of the most basic matrix, *to understand means next to nothing*; understanding is merely a translation of the new into the language of the old, of the already understood; a transcription into an old register, an adaptation.

If, for example, as a person without the slightest talent for mathematics, I have managed to understand a little about the backpropagation algorithm, this is not, of course, due to my "mathematical eyes" unexpectedly opening in my mature years, but it is purely due to the fact that, with the help of a multitude of good popular science articles, I have succeeded in bringing the idea of this algorithm within the parameters of a conceptual apparatus that is familiar to me. The example may be a bad one; there are undoubtedly forms of understanding—for example, scientific understanding inscribed in formulae, theories, and calculations—which are not simply a translation of the unknown into the language of the known. But in the "psychological" sense—that is to say, in the sense of the process that goes on in our minds when we "understand"—the definition is not inaccurate: in order to understand something new, we must in some way place that newness in the coordinates of what is already known, and in so doing we undoubtedly inflict some loss on the newness itself.

And therein lies, no doubt, the more sophisticated hidden core of Nietzsche's definition: the process of understanding *itself causes a certain loss*—what is supposed to be its goal, the understanding of the new, will at best *return* from some journey through the past, which will leave traces of old delusions on the new, if the latter is to be understood. Understanding is a process that adapts and therefore *falsifies*; and this, of course, should not surprise us with regard to any process that has evolved as a means of maintaining fundamental errors. Ultimately,

---

19 Nietzsche, *Joyful Science*, 224.

we could say that understanding is the form of thinking which is *the most primordial* of them all—in no other form of thinking does the original task of the "cognitive apparatus," falsification, adaptation to vital errors, manifest itself so markedly as in understanding. The question thus arises almost spontaneously: we are told that there is a good chance that AI will reach the stage of human understanding in some not necessarily distant future, but what *if it has already missed* that moment in its development from the outset? Shouldn't it, if human understanding really originates in a falsification, be approaching human understanding *regressively*, like a crab? Is it not AI's tragedy, if human understanding really counts for anything, that—even though it emerges from the pure nothing of absolute contingency—it is nonetheless born in a form *not sufficiently underdeveloped* to be able to "understand"? Understanding may not be "too late" after all, as Hinton suggests, but premature.

Nietzsche has a natural place in Heidegger's theory of the intra-philosophical event, the theory of *Geschehnis*; he does not need to be imposed on it. In two ways. We indicated at the beginning that Nietzsche appears when Heidegger raises the problem of the unfolding of the history of Being, following his introduction to *Geschehnis*; in particular, in relation to the current historical moment, which is, according to Heidegger, unique in that there is nothing going on with Being within it. If, at the level of a *Geschehnis*, Being as such, through the counter-question to the question "Why?" reveals itself in its *essential opacity*—for the counter-question is, rather than our own thought, *the act of its essence*, the expression of the "self-concealment" of being qua being—then it can be said that this primacy of the negative is also maintained at the level of the problem of the history of Being: since our era is characterized by the complete self-concealment of Being, any intellectual apprehension of Being must be preceded by a *genuine thought-experience of its nothingness*. The name for this experience of the evaporated Being is, in Heidegger's philosophy, Nietzsche.

But Nietzsche is also connected to a *Geschehnis* in another way, as an almost indisputable source of inspiration. It is impossible to overlook the profound similarity with Nietzsche's famous introduction of the problem of the *value of the will to truth*:

> The will to truth that will yet seduce us to many a risk, that famous truthfulness of which all philosophers so far have spoken with deference: what questions this

> will to truth has already laid before us! [. . .] That *we* for our part should also learn from this sphinx how to ask questions? *Who* is it, really, who asks questions of us here? *What* in us really wants "the truth"?—Indeed, we stopped for a long time before the question about the cause of this will—before we finally stopped completely before an even more thorough question. We asked about the *value* of this will. Suppose we want truth: why *not rather* untruth? And uncertainty? Even ignorance?—The problem of the value of truth stepped before us—or was it we who stepped before the problem? Who of us is here Oedipus? Who the sphinx? It is a rendezvous, so it seems of questions and question marks.—And can you believe it, it finally seems to us as if the problem had never even been posed before—as if it were seen, looked in the eye, *risked* by us for the first time.[20]

The questioning of philosophers—sceptics no less than dogmatists—has been driven for centuries by the will to truth. It is not easy to get off this train, even if we wanted to: a sceptic who fights against the pretensions of true knowledge is no less a fighter for his truth than a dogmatist.

This changes only when a new question comes before us—no doubt an expression of a particularly sharpened truthfulness—which causes the will to truth to fall into question, and with it ourselves, who, even with this new question, are still being driven by it. This is the question of the *value* of the will to truth. Against what can we measure this value? What makes this value questionable? Undoubtedly, *life* itself, which is based on principles that are opposed to truth: the will to appearance, deceit, and error.

In this, one need not necessarily see adversity to truth or its relativization; on the contrary: especially in the crucial years 1881–1882, when Nietzsche began to develop the theory of embodied errors, his thinking revolved around the very question of how to smuggle truth into a life dominated by errors; how to assert truth within our apparatus of thought, which is composed of the very traces of the vital errors of the ego, of the entity, of being, of permanence. In short, I would not say that in this measurement of the value of the will to truth alongside the "benefit for life," one should see the primacy of life over truth, the devaluation of truth. Rather, one could say that by asking the question of the val-

[20] Friedrich Nietzsche, *Beyond Good and Evil*, in *Beyond Good and Evil, On the Genealogy of Morality*, trans. Adrian Del Caro (Stanford: Stanford University Press, 2014), 5.

ue of the will to truth, we do indeed, as Nietzsche wrote, expose *ourselves* to a certain risk, but at the same time reintroduce truth as a philosophical problem. Truth is now no longer "merely" an ideal that shines from above, and to which we can at best put a prism that directs the light of truth into the dark corners of the world below, but an *enigma* that gives philosophers an opportunity to take it up again, perhaps even for the first time.

So, why does Nietzsche describe this line of questioning as a "meeting of questions and question marks"? Since the question of the value of truth is also undoubtedly guided by truthfulness, by the unconditional will to truth, the question is: Is it really we who have questioned the will to truth, or, on the contrary, has it been the will to truth itself that has led to both the question of truth and *the question of the philosopher*, who, for the sake of truthfulness, has found themself challenged as to the meaningfulness of their existence? There is no doubt, therefore, that within the question of the value of truth there is at work a kind of recoil, a *Rückstoß* of the question *from its content back towards itself*, which constitutes the main formal characteristic of a *Geschehnis*.

It is in this same sense that also Nietzsche's miniature definition of understanding, "to be able to express something new in the language of something old and familiar," needs to be explained. It does not aim at giving a definitive answer to the question of the essence of understanding: by internally linking understanding to its "opposite," i.e. falsification, this definition does just enough to make it possible to raise the question of the *value of understanding as such*—in other words, it creates this strange condition that makes it possible to answer the question of the meaning of understanding by saying "almost nothing."

However, despite this not very encouraging assessment, in this case too, and even more so than in the case of truth, raising the question of value does not necessarily mean the same as to devaluate. For, in this respect, understanding is a very special concept: that which seemingly deprives it of legitimacy—its connection to the fabrication—raises it above all other forms of thought in the scale of the value of life, since, by virtue of this connection, it turns out to be a life force, a condition for the survival of the species. On the other hand, understanding, of course, cannot be excluded from the value scale of truth either, since it represents one of the few accesses to it that cannot simply be abandoned.

This is the reason why a *Geschehnis* in Nietzsche's philosophy had to take place on the level of truth and not on the level of understanding. Actually, there are two reasons:

Firstly, understanding in itself does not have a natural antithesis similar to the one Nietzsche discovered in life regarding truth. Understanding is not the antithesis of falseness, with which it is intrinsically connected, it is not opposed to life, it is not opposed to truth; it does not possess that pure, strong opposite in relation to which *its value could be measured*. Or should we say: "*it had not had such an antithesis*"?

And secondly, we cannot simply abandon understanding—that would be like stepping out of our own skin. With regard to truth, even if we are reluctant to admit it, it is possible, if nothing else, to imagine its complete extinction: it is possible to imagine, as Nietzsche wrote, that truth itself will at some point turn out to be just another of the errors that served for a time as the conditions of the survival of the species, and then themselves were outlived. On the contrary, understanding—whether real or delusional—appears to be irreplaceable, part of our constitution. Or should we say that *it has shown itself to be irreplaceable*?

Both now exist: understanding is now confronted with its antithesis as well as with its nullity. Intelligent systems that learn even if they do not understand; intelligent systems that, without understanding, can take on the difficult tasks of understanding; intelligent systems that, even if they know nothing, produce real knowledge—undoubtedly, understanding now has both an antithesis and a rival, which is already gradually replacing it.

It is not necessary to put forward a speculative thesis as to the fact that the value of understanding is being questioned today; it is enough to call as a witness a certain fear that is spreading. It is not insignificant that Hinton himself has recently joined the ranks of those who call for caution. Nevertheless, I am almost certain that at least part of Nietzsche would have been sympathetic to intelligent machines: he would have thought that in these inanimate entities, entirely made up of errors, life as such is returning to its essence.

References

Heidegger, Martin. *Einführung in die Metaphysik*. Tübingen: Max Niemeyer Verlag, 1953. Translated to English by Gregory Fried and Richard Polt as *Introduction to Metaphsyics* (New Heaven: Yale University Press, 2000).

Hinton, Geoffrey. "The Godfather in Conversation: Why Geoffrey Hinton is Warried About the Future of AI." YouTube video, uploaded by University of Toronto, June 22, 2023, 46:20. https://youtu.be/-9cW4Gcn5WY.

———. "Two Paths to Intelligence." Lecture at University of Cambridge, May 25, 2023. YouTube video, uploaded by CSER Cambridge, June 5, 2023, 1:10:31. https://www.youtube.com/watch?v=rGgGOccMEiY.

———. "Will Digital Intelligence Replace Biological Intelligence?" Romanes Lecture at the Sheldonian Theatre, February 19, 2024. YouTube video, uploaded by University of Oxford, February 29, 2024. https://youtu.be/N1TEjTeQego.

Nietzsche, Friedrich. *Beyond Good and Evil*. In *Beyond Good and Evil, On the Genealogy of Morality*, translated by Adrian Del Caro, 1–203. Vol. 8 of *The Complete Works of Friedrich Nietzsche*. Stanford: Stanford University Press, 2014.

———. *Human, All Too Human (I): A Book for Free Spirits*. Translated by Gary Handwerk. Vol. 3 of *The Complete Works of Friedrich Nietzsche*. Stanford: Stanford University Press, 1995.

———. *The Joyful Science*. In *The Joyful Science, Idylls from Messina, Unpublished Fragments from the Period of the Joyful Science (Spring 1881-Summer 1882)*, translated by Adrian Del Caro, 3–285. Stanford: Stanford University Press, 2023.

———. *Nachgelassene Fragmente 1884–1885*. Edited by Giorgio Colli and Mazzino Montinari. Vol. 11 of *Sämtliche Werke: Kritische Studienausgabe*. Berlin: De Gruyter, 1999.

———. *The Will to Power*. Translated by Walter Kaufmann. New York: Vintage, 1968.

Sutskever, Ilya. "Why Next-Token Prediction is Enough for AGI." YouTube video, uploaded by Dwarkesh Patel, December 13, 2023, 2:07. https://youtu.be/YEUclZdj_Sc.

Tegmark, Max. *Life 3.0. Being Human in the Age of Artificial Intelligence*. London—New York: Penguin Books, 2018.