Lorenzo Perilli, Director of the Department of Literature, Philosophy and Art History at the University of Rome Tor Vergata, is a philologist and historian of ancient scientific thought. He is the director of the Interdisciplinary Study Centre on Classics, Mathematics, and Philosophy "Forms of Knowledge in the Ancient World" and has been involved for many years in the study of the digital environment. His first book on the subject, *Computational Philology* published by the Accademia Nazionale dei Lincei, dates back to 1995. He is the director of the journals 'Technai: An international journal for ancient science and technology' and 'Science and technology for cultural heritage'.

Italian Publishing House Il Saggiatore in 2025 published his book, *Artificial Consciousness. How Machines Think and Transform the Human Experience* where he discusses the problems that machines equipped with intelligence and deep learning capabilities pose us - problems of psychological adaptation, feelings of inadequacy, new ethical challenges, a new type of human-machine relationship, and technical-scientific issues such as the difficulty in understanding the deep mechanism and functioning of these machines.

Nineteenth-century terrors expressed in literature return: Frankenstein, robot rebellion?

We asked Lorenzo Perilli several questions on these issues. For some of them, we know there is no answer; what matters, however, is to be able to ask them.

https://www.ilsaggiatore.com/libro/coscienza-artificiale

Generative Artificial Intelligence: How have we come so far?

Lorenzo Perilli

Every day we read of new and amazing successes of Artificial Intelligence, which has entered our lives with products that are now almost commonplace. Some of us, especially those who have followed the developments of these technologies over the decades, not without some perplexity wonder: How did we get to generative AI? How was this sudden acceleration, these achievements, possible?

Actually, we don't really know how we got here. Throughout the 1980s and 1990s, and even beyond, people continued to talk about Artificial Intelligence in the terms in which it had been thought of in the 1950s. Everyone who dealt with it in those years realised that following the approach of John McCarthy and the AI lab at MIT in Boston or at Stanford would not really go far. And, in fact, for a good twenty years there were no particularly significant developments.

At some point between the 1990s and the year 2000, everything suddenly changed: another idea of artificial intelligence, no longer based on a symbolic approach, on the manipulation of symbols and logical rules, took the place of the original one. This corresponded to a return to the idea of neural networks, which had been designed many years before with the Perceptron without great results. Scientists developed new multilayer neural networks with many hidden layers, able to support what is now called Deep Learning, made possible thanks also to the increased computing power and the availability of very rich datasets for training the algorithms.

This switch, even according to experts, is difficult to control. Indeed, it is not controllable at all. Those working on AI tell us that the behaviour of deep neural networks is neither fully explainable nor transparently communicable. The widespread opinion - and the resulting enthusiasm - is due to the fact that the results have been far beyond the expectations. Chatbots such as ChatGPT or Claude or Gemini have shown that machines, following a change in the way we train them, learned to respond to our requests by speaking our own language, using our grammar and syntax without anyone teaching them – as computational linguistics had attempted to do in the 1980s. Joseph Weizenbaum's ELIZA, back in 1966, was an early example of a machine capable of interacting (by merely writing on a screen) with a human in natural language, albeit at a very basic level and with modest performance compared to what we see today. Despite its limitations, despite the fact that it used pre-organized sentences obtained using the same words as the user, ELIZA had a very powerful psychological effect on the individuals who entered into a dialogue with it. Today, however, machines have acquired the ability to learn on their own, to self-correct, and to autonomously intervene in problems that may arise even during interaction. Algorithmic machines today largely operate 'on their own'.

We are therefore witnessing a radical, and to some extent dramatic, change in the relationship between subject and object. We are not helped in our understanding of what is happening by analogies with the technoscientific revolutions of the past, the spread of electricity, the industrial revolutions, not even the advent of the Internet. Never before has the relationship between subject and object been reversed as it is today: machines becoming the active subject while we find ourselves as passive users when faced with the activity and decisions taken by machines. Machines to which we ourselves have delegated decision-making of any kind.

Could we ever have imagined talking to machines and receiving 'human' responses? Yet by now it seems natural to us, we command and control machines with our voices, an activity that produces on us, being unaware of it, an impact that is by no means taken for granted and is on the contrary profound, as Weizenbaum had already pointed out with ELIZA, when he wrote that 'Most people don't understand computers to even the slightest degree. Thus, they usually explain the computer's intellectual feats to themselves with the single analogy available to them, the model of their own capacity to think.'. People cannot understand how a computer works and consequently they associate it with something they already know, i.e. with their own psyche, human psychology, their way of behaving and speaking. This analogy implies that a machine endowed with such interaction capabilities effectively becomes a natural interlocutor.

In fact, one of the founding fathers of AI, Alan Turing, who had foreseen many of the upcoming developments, cleverly suggested in his famous 1950 *Mind* article, that in order to understand whether a machine can think, it is first necessary to define what we mean by thinking and what a machine is. Today, we tend to disregard a definition of these concepts and take them for granted. We are not able, or no longer want to define the concepts of intelligence, of machine, and in order to make up for these shortcomings we resort to the new totem, the word *algorithm*, again in absence of a definition. We have lost the ability to define the concepts we are talking about.

Human intelligence and artificial intelligence: just a question of scale?

We often hear that neural networks mimic the functioning of the biological neural networks of the human brain. That our brain is a kind of machine, a computer that performs calculations by running on electrical impulses. As a matter of fact, whereas machines need billions and billions of data to work with, our brain manages to gain a deep understanding of complex phenomena on the base of a limited experiential activity. In order to understand that, in a standard sentence, the definite or indefinite article cannot be the last element – an observation we arrive at after limited interaction with our environment –, the machine must have been trained on an enormous amount of textual data, so as to recognise probabilistically that an article at the end of a statement, not followed by a noun or something else, makes no sense. There is a profound conceptual mismatch between the way in which algorithms work, and the human way of thinking and acting, a radical, irremediable heterogeneity. The algorithm demands unambiguity, non-redundancy, effectiveness, efficiency, determinism, i.e. exactly the opposite of how the human mind works. Humans feed on ambiguity, human nature is intrinsically ambiguous.

The path that has been taken by employing neural networks as the basis of artificial intelligence tends to conceal the process of AI development, not least because, to date, the transparency and explainability of AI that everyone is calling for is still far from being achieved, and the deeper these networks are, the less comprehensible and explainable they become. The expected results are there, in front of us, but this is misleading, since they hide a process that keeps becoming more and more complex and for that very reason mysterious and incomprehensible as well as less and less explainable and justifiable.

We humans hardly ever *go beyond ourselves*, 'renounce' our world, and we rather transfer our models into our representation of these fascinating machines: this is inevitable. Hence comes the comparison artificial brain-human brain, but also the parallel between human ethics and artificial ethics. We attribute to these machines union rights, features that are typical of consciousness, autonomous decision-making capabilities as we would to ourselves. In reality, we should completely reappraise our relationship with these machines and try to understand them in the specific reality they belong to.

The human is not univocal, it does not live on efficiency, but rather thrives on ambiguity and error. If this machine revolution is accepted and embraced without critical judgement, we will be reconfiguring the very nature of the human, claiming to characterise it on the basis of the parameters of algorithmic machines: univocity, efficiency, progress in the shortest possible time. If that is the case, the machines have already won: it is not even worth playing the game.

Our being human undoubtedly lies in doing; throughout history, the human species has always produced tools, techniques, and technologies to overcome the limits and constraints of the surrounding nature – otherwise, we would all have died long ago, as we are a weak species within the natural world. However, this has of course happened in order to achieve practical results, but in doing so, in the course of these processes, we have learnt from successes as well as from failures, and we have grown cognitively. If we abdicate this slow process of learning and comprehension, at a time when intelligent machines can perform higher cognitive tasks better than us such as writing, speaking, doing mathematics, creating images, writing code, while we not only do not understand how they function, but we also give up understanding and ask the machines to understand and explain themselves to us, what can we predict will happen to the development of our own capabilities?

We are on our way to a new world; of this I have no doubt.

I've been working in the digital field for many years, but I never cease to be amazed when I read about the results of stimulating experiments. In a lab at Duke University in the United States, electrodes connected to neurons were implanted in the brains of small rodents in a cage—a neural implant not unlike the ones now being discussed for humans. Thanks to this device, the mouse in the cage, merely by feeling the urge to drink, transmits the corresponding signal to a computer outside the cage, which then opens the tap. Now that's what you call the power of thought. Of course, with humans it will be more complicated, but we are now capable – and not just theoretically – of restoring certain abilities to people with disabilities, particularly those paralyzed because of severe spinal cord injuries: in laboratories in Lausanne, Switzerland, there are patients who can walk by controlling movement with their thoughts and bypassing physical injury. We're seeing the same thing with truly optimistic prospects for smart prosthetics, or in robotic surgery, which is now a widespread reality. I'm convinced that it won't be long before the surgeon will no longer be operating at the console, but rather an artificial intelligence system that autonomously controls the robotic arm, with the doctor merely supervising.

Another area of research, that of affective computing, makes it possible to read and interpret emotions on people's faces—often recognizing them better than we would ourselves. But that's not all: these machines are capable of simulating human emotions and feelings through voice and performance. Forget the Turing test. How can we think this won't have a huge impact on us, on a child growing up interacting with such systems?

We look at these results in amazement, and are astounded by the ability of the new machines to interpret vast volumes of data, and at the same time, we are confused and disoriented by the fact that we cannot intervene and get involved, because we don't truly understand what is happening within those processes. I don't believe that machines will take over the planet or come to dominate us - I don't believe in catastrophism of this kind. However, I do believe—and I'm convinced—that we will gradually adapt our concepts to those of the machines; we will adapt to a mindset of efficiency, where what matters is no longer the process that leads to a result—as it has always been for humanity—but only the result itself. And if that's what truly matters, then machines will always be better and more efficient than humans, because we, as machines, are fundamentally imperfect. But it is precisely this imperfection that makes us unique, and therefore, interesting.

But can there be an ethics of artificial intelligence?

This question is actually more of a provocation, but it raises a real and deeply felt issue, even if one that's often overused to the point of becoming a stereotype.

If we consider the many and diverse ethical theories, if we contemplate our own personal ethics - the one each of us feels we possess - or the ethics taught by religions, we'll see that ethical concern has never truly entered human history. Simone Weil expressed this quite clearly, and we see it every day: human history does not operate in accordance with ethics. We human beings have created beautiful structures – art, music, science, even ethics itself – but all these extraordinary creations, in truth, lie outside of history. They pass through it in specific moments – brief, lightning-fast moments - and when this happens, we are amazed: we are amazed precisely because they are exceptions.

Let us pause for a moment to think about smart weapons and military operations, which are already largely based on AI – and unfortunately, this is the reality of recent months.

The identification of specific targets is now often carried out using AI systems, which in a matter of seconds select and detect the location of thousands of human targets, based on data collected – data that the targets themselves have entered into their profiles – gathered by linking together various so-called metadata, that is, the traces we leave while moving online, or by digging through the contacts in digital address books, daily habits, data from phone surveillance, frequently visited stores, purchases made, websites visited: by combining and analysing these and other data, an algorithm classifies the target and pinpoints its location. The machines decide who will live and who will die, in a decision-making time of just a few seconds, at most asking a soldier to validate the choice—giving them no more than twenty seconds to do so.

Possible errors? Certainly. But they are accounted for. If the target is low-level, a limited number of what are called "collateral victims" is considered acceptable – up to ten. If the target is high-level, the number of acceptable collateral victims can rise to one hundred. An entire building may be bombed, regardless of who is inside, if this ensures that the high-value target is eliminated.

Where is the ethics in this? In accepting a balance between collateral victims and the threat level of the target? One often hears calls for an ethics of artificial intelligence, when what is primarily needed is an ethics of human action.

Norbert Wiener, the most aware among the great scientists of the twentieth century, never tired of warning us: these machines do not share our principles, nor do they make decisions based on our emotions. Despite everything that had happened during the 1930s and 1940s, Wiener still wanted to retain a certain faith in humanity. But even assuming that these much-vaunted human principles truly exist, we should ask: isn't it perhaps an advantage that machines follow different principles than ours? Some argue it is – for instance, robot soldiers would not act under the influence of emotions that might compromise their decisions. Yet this marks an extraordinary phase transition: it means abdicating our humanity, for better or worse; it means erasing 2,500 years of ethical reflection, from Socrates – indeed, from the heroes of the *Iliad* – to the present day. One might argue that our ethics, the same ethics Aristotle sought to formalize in the *Nicomachean Ethics,* have only ever entered human history in exceptional moments. Even the much-praised International Declarations of Human Rights merely stare back at us from elegantly printed pages, but they have never truly been shared, and even when they are, they are only applied when deemed harmless.

We must place this discussion in the present, in the here and now, in a world dominated by a radicalized form of so-called Western capitalism – and this is merely an observation, with no political connotation – a world and a form of radical capitalism that has encouraged behaviours and actions aligned with the "philosophy of machines": the idea that what truly matters is the achievement of the goal, of some advantage – whatever it may be – generally equated with profit.

And so, I provocatively ask myself whether an "artificial ethics" can truly exist, and whether that is good or bad—and, if it does exist, whether it is good or bad that it be, in any case, influenced by us and by what we like to call our own ethics.

An ethics of duties

We must not make the mistake, one we have made many times throughout history, of underestimating the deepest traits of human beings – traits shared with other animals – described by the ancients with the concise yet powerful expression *homo homini lupus*. The history of ethics in societies, along with legal, social, political, and religious rules, is the history of an attempt to moderate and sanction this tendency toward domination over others. In many cases, these rules have worked and still work: if we look at what we have achieved, we can say that yes, several steps forward have been made. But that underlying trait has not changed, because it is rooted in biology.

Anthropologists, psychologists, and psychiatrists explain that our behavior follows the principle of the in-group and out-group – that is, being part of a group or not – and they emphasize how this affects our actions. There is a biological foundation that governs our behavior: being part of a group, feeling at ease, recognizing oneself in it, and accepting its rules means being protected, being defended from those outside the group who might want to join it. Others are different simply because they are not part of the group: skin color, culture, religion, the school attended, the football team supported. This biological element shapes each individual's inclination to accept the rules of their own group, and not those of others. Hence the difficulty – if not the impossibility – of implementing an ethics based on moral principles shared by all. Simone Weil wrote this clearly at the beginning of the 20th century, in her book *L'enracinement, prélude à une déclaration des devoirs envers l'être humain* (*The need for roots: prelude towards a declaration of duties towards mankind*), when she called for replacing the word "rights" with the word "duties," and thus suggested that we should speak not of the rights of man, but rather of duties toward human beings.

Rights, Weil said, only acquire value if they are recognized by someone else or by an authority that defends them, whereas the duty toward others is something we feel within ourselves. No one needs to grant it to us, and it would still be valid even if we were alone in the universe. When the French Revolution introduced the Declaration of the Rights of Man and specified, however, that they were the "rights of man and of the citizen," it created a contradiction: on the one hand, it spoke of universal rights owed simply by virtue of being human; on the other, it referred to the rights of someone who belongs to a State, a member of a group defending its own prerogatives – even by force. Universal human rights then become secondary to those of the citizen: if I identify with and claim certain rights as a citizen – that is, as a member of a group – those rights will eventually prevail over broader rights that could endanger that group.

The commitment to a shared ethic is commendable, but we cannot ignore that the difficulties in achieving it rest on deep biological and anthropological foundations, because the moment we see ourselves as members, we exclude others – and this already creates the conditions for potential conflict. We must ask ourselves why, throughout history, there has never been a population that did not develop at least some form of religion or religiosity – perhaps as a way to provide a foundation for this other undeniable need of ours: the need for connection with others. This metaphysical need, for something beyond ourselves, has been the source of those extraordinary and brilliant moments of humanity in which beauty prevailed over profit and selfishness, and which gave humanity moral norms to follow and the hope of being able to do so. Religion thus served the purpose of imagining an external authority, different from us, untouched by our biological characteristics, capable of guiding us toward different actions. That religion has also been and can still be abused – this too is history.

It is not easy, therefore, to give up mental structures that are biological and deep-rooted. It is even more difficult to extend a shared ethic to machines, to the various types of intelligent machines used around the world. In this context, what we call ethical guidelines are in fact reduced to rules of use, laws, protocols. But ethics cannot be imposed through legal means. At most, through culture – but that takes effort. Ethics is something different, or at least it should be.

Actually, the world of machines presents itself as an extreme example of group behavior, because the machine has been designed and programmed to do better, to do more, to prevail, to achieve the result in the shortest possible time without concern for the process or the path taken. The most trivial example, which anyone might experience daily, is enough. The doorbell rings: I get up from my chair and walk 10 meters to the door, perhaps go down some stairs, turn on the light, open the door. In this whole process, my brain has developed, my sensitivity has had an experience, I have grown. If I tell the smart home system, which responds to voice commands, "open the door," I have already lost the entire process, I have turned inward, I have renounced my relationship with the surrounding environment. I do not look at it, it is not useful, I do not need it.

This is the new world we are entering, a world in which we will no longer need others nor to interact with them, a world that requires completely different premises from those to which we, as social animals, have been accustomed so far. Above all, we are required to make at least the effort to become aware of the world we are building, because artificial intelligence is, in turn, an extraordinary result of our own intelligence. It is humanity's oldest dream – the dream of automatic, powerful, intelligent helpers, similar to us. The dream of the god Hephaestus, the blacksmith of the gods, in Homer's *Iliad*. The only god with a trade. After 2,500 years, that dream has come true. Now we should be careful not to overdo it.

Indeed, artificial ethics – the ethics of machines or of the use of machines – is nothing other than our own ethics; it is the mirror of what we want to become as human beings. Where do we want to go? What are we doing? What goal are we trying to pursue? Do we really need all this efficiency? We have created the most serious problems ourselves; we created the climate crisis, we constantly create new diseases, and then we strive with our intelligence to find ways to address these very problems. With new algorithmic machines, we have truly reached the point of having solutions in search of problems.

To describe this paradox, the philosopher of science Paolo Rossi once used the myth of Daedalus: he who invented the labyrinth and then had to draw the map to escape it.

Here we are.