ISSN: 2282-4189

LUISS

Research Center
for European Analysis
and Policy

EMUNA Brief 3/2026

SCIENCE AND ETHICS IN DIFFERENT CULTURES – 9 February 2026

Mario Rasetti

AI: Work in Progress. The Challenge to Thinking Machines: Intelligence, Ethics, Truth

AI: Work in Progress

The Challenge to Thinking Machines: Intelligence, Ethics, Truth

Mario Rasetti[1]

Abstract

A concise overview of the most recent advances in Artificial Intelligence, with special focus on the ethical issues and on strengths and weaknesses, and on how these bear on the possibility that there might emerge risks for human beings.

[1] Politecnico di Torino.

The world we live in is inexorably characterized by two factors: 1) the great instability of the reference framework (geopolitics, environment, resources, economics, etc.), which examined in light of the data, reveals the system to be a hyperbolic dynamic system: like a tightrope walker, all its points of equilibrium are unstable; sustainability and stability are the issue; 2) the fact that everything is grafted onto a vast, rapidly growing digital multi-platform, which we must master.

At the center of this framework lies Artificial Intelligence (AI), which is—as always happens with great scientific revolutions—both angel and demon: on the one hand, a tool that can make humans more powerful and efficient, providing them with unexpected means to multiply their capabilities (of the "head," not of the arms), on the other, a "monster," for example a financial one.

AI has extraordinary strengths, but also severe weaknesses. First of all, it has the wrong name—it does not have much to do with human intelligence—and a distorted image, tainted by too much science fiction, and it is experiencing a dizzying growth dynamic, with an unimaginable scope and width. AI is certainly the greatest cultural and (perhaps) anthropological revolution in the history of mankind, in which it will play as well a crucial role as catalyst for a new evolutionary leap. It could also be said that AI is inevitable: digital technology, the gigantic data production it entails, and the need to extract something useful and efficient from it are now part of the lives of all of us, while AI is the only possible tool that allows us to do so. AI is a superb technology, but it's not yet a science, because we still don't really know what intelligence is and how it is generated, whether by the human brain or by a machine. A first real challenge is to make of AI a Science.

AI has significant strengths: it better manages the complexity of information, has great brute force, and its branches and specializations (Low, Strong, Generative, General, ...) are growing and maturing ever more rapidly. But AI also has numerous, complex weaknesses that cause debilitating limitations: i) Energy (AI already uses over 4% of the electrical energy produced on the entire planet; a human brain requires 20 W of power); ii) Ethics, knowing how to distinguish good from evil – which affect democracy, the role of the individual, social values, truthfulness, collective intelligence, etc.; iii) it still is not able to resolve fundamental questions such as Causality, Learnability, Foresight, Truthfulness, etc., which – incidentally – all converge on Ethics: without them, we cannot promote it.

The applications of AI are now innumerable: from personalized medicine to tree atlases, to the relationship between protein sequences and three-dimensional structures, to personalized pharmacology; from industrial and commercial strategies to risk prevention; from the design of multimodal systems for generating and building logical structures, rational thought, codes capable of self-adapting or producing new codes, similar to personal AI assistants, to the simulation of complex systems, to 'scientist AI'; ...

A good metaphor to start with comes from the so-called LLMs (Extended Language Models) now evolving into LTMs (Extended Reasoning Models): beyond the GPT, a possible future for AI as a reasoning machine is increasingly visible. Generative AI, however, is not yet General. The path to get there is much longer and most likely a dead-end, because it requires computation with new rules of computation that go beyond Turing and the management of undecidability. But we already know that an LLM, which communicates with another LLM

for the science of complex systems is the source of an "emergence", which can measure the behavior of a human interlocutor participating in their dialogue: the first step toward understanding self-awareness, the predictive brain, the mind, etc.

Machine learning is learning, yes, but first we must unravel the knot of "understanding (comprehension) vs. knowing": this is perhaps a glimmer of General AI. It gives us hope that we will possibly be able to "align ourselves," that is, to distinguish good from evil (and thus solve Kurt Gödel's problem of free will), but understanding more about the mystery of an evolutionary pattern that could be summarized as: "Intelligence = Life $\oplus$ behavior $\equiv$ Complexity (emergence)" remains beyond our strength.

A frequent question is: should we fear the emergence of a superintelligence whereby machine creativity will render human intuition irrelevant? I think not, because we are capable of emotions and feelings; because our brain is not only the prefrontal cortex, but also the amygdala, mirror neurons ... and much more; because we know how to promote the improbable and shift paradigms ... all things that the homo sapiens-sapiens-sapiens that AI will help us evolve will increasingly be able to do more and better than their intelligent machines.

Nowadays, I'm afraid of human beings far more than I fear potential super-intelligent machines!

Let's review some of these points in greater detail and critically, to understand how to equip ourselves to face the future.

Representing our planet as a unique, global, complex dynamic system, generated by the interweaving of such diverse elements as geopolitical, socioeconomic, environmental, climatic, etc. that characterize it, it is possible, thanks to the extraordinary tools of HPC (High Performance Computing; top-power computation!), to simulate its dynamics. We immediately realize that this system is "hyperbolic": all its equilibrium states are in fact unstable equilibria. This implies that even the slightest perturbation can lead to disaster: the Earth is like a tightrope and humanity is like the tightrope walker walking on it, exposed to every breath of wind, which in a precarious balance like this, has the strength of a tsunami. The system is complex, the interactions within it are non-linear, thus control cannot help but be equally complex. With AI, the challenge is to do it as well as, and better than, what we could do with our brains alone.

In this context, digital technology has completely pervaded the lives of three-quarters of human beings—six billion out of eight and a half billion people own at least one cell phone or other device with which to access the internet—and in 2025, we have generated, distributed, and manipulated an amount of data equal to $7.5 \cdot 10^{23}$ bytes (seven hundred and fifty trillion gigabytes)—with a doubling time that, due to the simultaneous growth of the Internet of Things: machines that use and generate enormous amounts of data and are themselves actors and agents in the network (over 15 billion are expected to be activated in the next three years), is already less than 24 hours.

In this scenario, from such perspective, AI is inevitable. In a hyperbolic world, an extremely rapid and powerful control system is needed to cope with such extreme, unprecedented characteristics: only with AI this can be done, even if from time to time it makes mistakes (for good luck we call them hallucinations, but—as Alan

Turing taught us—"If a machine is expected to be infallible, it cannot also be intelligent,"). The only way we have to manage this dramatic, otherwise virtually uncontrollable global scenario—with an unknowable reference reality, of which, as in Plato's myth, we perceive only the shadows it casts on the bottom of the cave; representations that we can construct only thanks to the immense amount of digital data available—is AI, because fortunately we now have the computational power necessary to activate it, even if perhaps, paradoxically, we risk not having the electrical power to do so.

AI operates at various levels: from the "low" (expert systems), to the "strong" (optimal decision-making processes and strategies), to the "generative" (chatbots, machines capable of to speak with humans the same way they speak to each other, but with essentially unlimited knowledge), to the "general" (the presumably hopeless challenge of developing a machine capable of emulating the human brain).

The problems that AI allows us to address are increasingly complex. It can be said without fear of contradiction that it is the greatest cultural revolution (even comparable to those brought about by the discovery of writing or of movable-type printing) and (perhaps) anthropological (the evolution towards the successor of Homo sapiens-sapiens?). A quantum leap in the history of the Homo species.

The key word here is obviously Intelligence. It's disturbing that we still don't have a rigorous definition or a way to measure what intelligence is. This is why I think AI is misnamed. In this, Alan Turing misled us a bit with his Imitation Game: the problem is not understanding how to distinguish the intelligence of a machine from that of a human in the case of a machine capable of doing things that humans do "with their heads" rather than "with their hands": it neither answers questions nor provides suggestions for how to proceed.

For humans, intelligence is primarily related to the brain. It is interesting to note how, throughout its evolutionary process, the brain has undergone progressive selection as an organ increasingly responsible for the protection and defense of the organism to which it belongs; this is why its evolution has both a collective (species-specific) component and an individual-specific one (precision intelligence, the mind). Culturally, we tend to think of intelligence as an absolute category and limit its evolution to the goal of identifying the origins of rational thought, but this often does not coincide with its primary function of bodyguard that the brain has.

The brain is the most extraordinary machine in the known universe; however, it is, now and perhaps forever, beyond any possibility of simulation or reproduction on any medium other than the biological one. Moreover, the brain's functions are not limited to learning alone; the brain performs many other functions (of many of which we are still unaware): the complexity of the connectome is truly immense, so much so that it seems almost infinite.

For now, we must limit ourselves—and that's already a titanic undertaking!—to saying this.

The strengths of AI, "thinking machines," beyond their powerful array of skills, are simple to describe. They allow us to better manage any form of complexity, discovering ever deeper correlations between seemingly unrelated information and using it to acquire (learn) a growing range of extraordinary capabilities, such as

prediction. This is especially true because AI possesses remarkable raw power, both in terms of processing power and memory (the ability to store and preserve what it learns in an organized manner).

The weaknesses and limitations, however, are numerous. First of all, the energy it requires: there will be no real progress in LLMs—at least if the approach they are currently being developed continues—without investing enormous capital in energy, with solutions that clash with the interests of sustainability and the well-being of the planet.

A second point is both structural and foundational. In humans, aspects related to feelings and emotions contribute significantly to the processes of intelligence. Now, while rational thought primarily and largely activates the network of neurons called the prefrontal cortex, the neurons at the core of feelings and emotions engage a different part of the brain: the amygdala. This observation is crucial. The entire history of AI is based on the fundamental intuition of Warren McCulloch and Walter Pitts, who, analyzing the results of a set of subtle experiments by Francis Crick, discovered how the functioning of the prefrontal cortex neuronal network could be naturally described in the formal language of George Boole's logical algebra: just three operations; AND, OR, and NOT (with which we know how to build the ubiquitous NAND, which is used to construct all functions made from strings of 0s and 1s). Then and there the digital age was born—which is now the key backbone ingredient of our lives—and with it the notion of artificial neural networks, circuits so close in behavior to the cerebral cortex that they can simulate quite well many of its properties. But for these highly efficient objects, we have neglected the amygdala, which unfortunately cannot be represented in Boolean form.

Then come the other open ethical problems: Moral (respect for users and their dignity: democracy, the role of the individual, social values, etc. are put at risk); and, gradually and subtly, Causation (the ability to recognize what the cause—or effect—of what: most human knowledge is causal cognition; almost nothing a machine learns is causal unless the examples used to train the machines explicitly identify and state the intrinsic cause-effect relationships); Learnability (criteria for optimally selecting the "training set," the set of data used for machine learning (ML) training, to ensure that the machine is capable of achieving effective cognitive gain; cases have been observed where the choice of training set in ML can lead to scenarios that are undecidable according to Gödel—Shai Ben-David); Truth (the most fragile; perhaps we don't even know what it really is).

Indeed, truth deserves a concise examination on its own of its philosophical aspects. Recall that the Latin word for truth, "veritas," derives from the Sanskrit "vrāta," which denotes a "fact," an "occurrence": it seems to suggest that only a fact directly perceived can be tagged as true, not its representations. The pre-medieval English word trēowe, from which "truth" derives, instead, indicated a moral value (for example, of the knights: "steadfast," "loyal," "reliable"). In fact, it is ambiguities like these—what do we mean, in the full rigor that the scientific approach requires, by intelligence, or by truth—that lead us to say that AI is not yet a science, but a practice, albeit a highly effective one, carried out primarily by scientists, because it can best be achieved using scientific tools and methods. A true science has "no-go" theorems and criteria for identifying optimal training (measurement) sets, which ensure we don't find ourselves in a gödelianly undecidable situation; as well as an absolute method for deciding whether a correlation between events is or is not a cause-effect relationship (causality).

An incorrect choice of the subset of data intended for training can lead to undecidability, that is, that vagueness of the life of the "observables" at the heart of Luigi Pirandello's "One, No One, One Hundred Thousand" which so well symbolizes the challenge of distinguishing the true from the false (good from evil): each of us has his own exclusive representation of the truth and these representations are so numerous (metaphorically one hundred thousand) that it is as if there were none, but we firmly believe that there is only one... In the scientific world, this echoes Marvey Minsky's statement: "Artificial intelligence deals with 'observables' that are all either ambiguous or undecidable, such as life, truth, intelligence, behavior, understanding, prediction, cause/effect relationships, the human being, ..."

The path to transforming AI into a science will be long and arduous.

LLMs demonstrate this. The specific case of GPT (whose active brain is a "Transformer," T, Pre-trained, P, and Generative, G) is highly significant in the AI landscape. The future of LLMs is uncertain (we have already mentioned that they must first overcome the energy barrier and not crash into it due to unsustainable solutions such as computational power) and they are still far from becoming General AI.

Above all, however, their mission is to address the challenge of moving in the direction of a new, more powerful mathematics, "beyond Turing." It is this need to overcome the computability barrier due not only to immense algorithmic complexity but also to the Gödelian decidability of the solutions we believe we find, which raises the key question: will we ever be able to predict human behavior and model self-awareness? (self-awareness: knowing one's existence); or a farsighted brain and mind, like those of humans?

The path is arduous (though perhaps not as much as one might fear) due to the unprecedented number of possible evolutionary outcomes it requires. We must learn to build intelligent machines capable of distinguishing—but in a form that could be universally shared—good from evil, with a systemic completeness that would one day allow them to have free will, capable of understanding (comprehension, not just learning and knowing), and capable of creativity and ingenuity comparable to those of human beings.

The obstacle lies in the problem of self-referentiality, which AI cannot address. An AI that improves itself recursively faces a logical dilemma; it must evaluate the reliability of its own reasoning, but can only perform evaluation processes that are themselves embedded in the system it must verify.

This leads to a fundamental asymmetry: i) if AI trusts its own reasoning, it risks generating blind spots; ii) if it questions its own reasoning, it will be unable to justify its next step. This brings us back, in a different formulation but with the same logical structure and meaning, precisely to Gödel's second theorem: an AI is unable to internally justify its own correctness. Any certification or assurance of safety must come from an external system—for example, human carelessness, or an external analysis—which in turn is finite and subject to its own gödelian limitations.

DRL is nothing but a cascade of n ML operations using methods, specifically convolutional artificial neural networks (ANNs), that do not address uncertainty in conventional probability approaches. Indeed, they do not

incorporate any explicit representation of the environment in which they operate, but the architecture of their networks, as well as the representation of their reference reality (the shadows at the bottom of the cave in Plato's myth) are also left free to evolve autonomously in the iteration process typical of DL (auto-regression), thus creating a somewhat 'specific' one at each step. This means that, once an ANN's training is complete, the programmer has no idea what calculations the network performed or why they worked. And if the network fails, he or she has no idea how to fix it. The process has become an indecipherable black box; the Black Box. Here is the dilemma: Bayesian networks cannot «understand» cause-effect: by design they are such that information flows in both directions, causal & diagnostic, and we need a different statistics frame. Yet Bayesian networks need a stable 'reference reality' and the Black Box changes its own at every step of DRL.

The problem is that, just like the prisoners dreamed of by Plato, the DRL explores only the shadows on the cave wall and learns to predict their movements, but ignores the fact that the observed shadows are simply projections of something that is itself a representation of a reality which is changed on the next step, absolutely different from the 'real' one. In other words, for the DRL process, reality sees a world that is nothing more than a representation of an (incomplete) representation of a representation of a representation … many times: a world "n steps from reality".

AI is therefore very fragile, and analyzing its weaknesses, we easily observe, somewhat surprisingly, that most of them are ethical in nature: 1.) Lack of transparency: its decisions are not always comprehensible to the humans who make them. 2.) Lack of neutrality: AI-based decisions are often (too often) imprecise, discriminatory, and influenced by biases—either intrinsic to the machine or learned from humans during training. 3.) Lack of control: surveillance practices in the data collection process or active user protection are often insufficient. 4.) High-level AI consumes a lot of energy: thus, it has a significant impact on the environment and the geopolitical landscape (e.g., water, rare earth materials, higher environment temperature, etc.). 5.) Limited scope: The brain is capable of performing a myriad of tasks that cannot be ascribed to learning; AI—at least for now—cannot.

Let's critically review (at least point out) them: 1) Freedom and Autonomy: We know that "free will" is possible if the system is quantum (due to the principle of superposition of states); however, we still don't know much about how a machine can have the 'courage of morality', the vision of good in a global sense, the complexity of thought contextualized within society. 2) Respect for Dignity: AI does not yet have the necessary and sufficient tools to construct this behavioral paradigm, because this is primarily a problem for the amygdala, not just the prefrontal cortex.

We need to be able to feel compassion (in the etymological sense of the term: sharing suffering), empathy, and a sense of collective unity. McCulloch and Pitts's choice of the computer as a metaphor for the functioning of the brain was unfortunately used backwards, thinking of the brain—the entire brain, not just the prefrontal cortex—as a metaphor for a computer, a Turing machine. Turing isn't enough; we need to go further (Gandy?). This is because biological memory is very rich; but—we now know—it is equally unreliable because it is incredibly interconnected: it can be accessed through multiple pathways and not just through a single address. Our brain may be similar to a computer in how it processes information, but our way of storing and, above all, recalling memories—on which the concepts of dignity and solidarity (between species) rest—is entirely

different. We humans are not machines, and if we are, we are not like any machine we have built in the past or are now capable of conceiving.

LLMs represent enormous progress for AI, but above all for human society and its organization, because they address language not only as an object in itself but as a communication system (chatbot) at a cultural level. GPT is certainly not yet a GAI, but it represents a significant step toward the (visionary/prospective) proof of its existence. Its choices simply maximize the probability of certain behaviors: we must not forget that creativity, on the other hand, promotes the improbable; genius breaks paradigms... their probability is low, but its effect can be disruptive. While ergodicity is "extension," it favors the growth of entropy and generates disorder.

We will probably never have a (super)intelligence—whatever that means and indicates—capable of governing the world and ourselves without us. Our most powerful weapon to prevent this from happening is mathematical proof. Gödel's Completeness Theorem (note: not the Incompleteness Theorem) states that any statement that is true in all models has a proof, which implies that if a property doesn't have a formal proof, then there must be a way to break it ... and a true GAI with sufficient power will surely find a way to do so. And if that GAI is hostile or controlled by someone with malicious intent, then that someone will exploit that weakness. Notice that none of this absolutely implies that Turing is right in stating that we should now "expect machines to take over."

NO: no matter how intelligent a GAI (even a "super") may be, it will never be able to do what is demonstrably impossible.