MARCH 19th , 2025

By Eric Stefanello,
Polytechicien, specializing in AI and the security of complex software and systems.

IA SAFETY: AN IMPERATIVE IN THE FACE OF THE UNKNOWN

In recent years, various scientists and engineers have begun to talk about the problems that neural AIs pose in terms of personal safety. This issue is becoming more pressing with the arrival of AIs interacting directly in and with the physical world.

Although they are bombarded with information about AI on a daily basis, citizens do not have a clear understanding of what these machines really are, even though they are being deployed more and more widely in our companies and in our lives.

The way they work represents a total epistemic paradigm shift because it uses the inductive approach rather than the hypothetico-deductive approach.

Neural AIs use purely inductive approaches. As such, they have no underlying representation of the world, whatever the subject they are working on. They are not built on the basis of symbolic representations of the world but on the basis of a massive statistical approach to the correlation of large masses of data.

Rather than artificial intelligence, they should therefore be called "massive data correlators".

This is how in LLMs (*Large Languages Models*) language, words are represented by vectors of 300 dimensions. Each of these words/vector has a meaning that is not at all symbolic but is based on the statistical proximity to the other numbers of other words with a similar or related meaning.

When this vectorization of language is well done, you get a representation that can be computed in the mathematical sense. So if you take the king vector and you subtract the male vector from it and add the female vector to it, then you get a vector that is extremely close to the queen vector. Extremely close to the point that there is no other vector as close to the queen vector. So you can infer that: king - man + woman = queen.

This computability is THE pillar of the operation of all LLMs. It allows you to make calculations on words, to sort them, to distinguish them, to classify them, to contextualize them, to interpret them.

This is where the learning phase comes in: the neural network will be made to ingest masses of data that will allow it to gradually correlate them and be able to infer from a series of computable words/vectors which is the next word that can most likely occur.

These neural networks are built by providing them with parameters that will intervene in each input and output of a neuron. It is learning that will allow the machine to gradually adjust each of these parameters until its performance corresponds to what is expected.

The program is not mastered or designed by humans, it is the machine that programs itself thanks to the data provided to it.

The "program" is the result of learning, not that of human work. Corollary: we don't really know what's inside AIs, nor do we control their behavior, and above all their behavior is non-deterministic: it can neither be predicted nor analyzed precisely a posteriori.

So, if there is intelligence, one day, perhaps, it will be totally, radically and definitively different from human intelligence. Neural AIs will probably be omniscient but will lack the forms of human intelligence that are specific to our biology: relational, psychomotor, musical intelligences and especially self-awareness, since the latter is linked to consciousness.

It emerges that AIs pose radically new security problems: will we ever be able to trust these machines, to the point of letting them pilot excavators in the street, planes full of passengers, water treatment plants, electricity grids, etc.?

With neural networks, humans have invented programs that "self-modify" and learn on their own, without our help except for our data or the data they acquire otherwise. We know how they learn, but we don't know the content of all their new learning. It is only when AI acts that its new capabilities are discovered. These are programs that are self-designed in a way, baby programs that become adults autonomously, through self-learning.

In terms of foresight, we can safely announce the massive arrival of AI in the real world in the next 10 years. These AIs acting directly in the real world will pose radically new security problems.

Indeed, the most complex software systems used in physical systems are until now totally deterministic in their behavior: piloting aircraft, controlling nuclear power plants, etc.

No computer system of this type has ever had a massive outage. The proliferation of computers and the certification of the software that runs on them creates a "safety by design" approach.

In the world of neural AIs of today and tomorrow, none of that. We are faced with programs whose behavior intrinsically conceals an indeterminate aspect, if only because learning constantly modifies the "program".

Another major risk factor is that while the mathematics that makes it possible to make these neural networks is simple, we do not have the mathematics that allows us to describe or understand how they work. Yes, you read that right: the mathematics describing how AIs work does not yet exist.

It should be noted that the computer cathedrals that are these large LLMs like Chat GPT with 1800 trillion parameters and more than 150 layers of neurons are more a geek and hacker's construction than an organized ordering of master masons under the direction of a wise architect. These buildings were built little by little with a game of trial and error with one constant: the results obtained always went beyond initial expectations and surprised the designers.

However, there is no possibility of controlling AIs if we do not have precise mathematical models of them. This question is fundamental.

Over the past 18 months, the creation of state bodies dedicated to the safety of AI has been multiplying around the world. The USA in November 2023, then Japan, Korea, Singapore, Canada, China, Argentina in 2024. The EU created a dedicated office at the end of 2024 and France has just created the INESIA (National Institute for the Evaluation and Security of Artificial Intelligence) which is not a dedicated entity but a federation under the aegis of the SGDSN, the various organizations involved in AI security and safety issues.

At the AI Security Summit in November 2023 in Bletchley Park and at the initiative of the UK government, a panel of 96 international experts from 30 countries was formed, as well as representatives of international organisations such as the United Nations, the European Union and the Organisation for Economic Co-operation and Development (OECD). The secretariat of the panel is provided by the British government, and Professor Yoshua Bengio will chair the panel for the year 2025.

It could be considered a kind of "IPCC of AI".

This panel has just released its report: "International AI safety report" available online in January 2025. After a few reminders on how AI works, this report, which has gone largely unnoticed in the press in recent weeks, despite the AI summit in Paris, provides edifying elements in terms of analysing the risks associated with AI.

Noting that the performance of the latest versions of ChatGPT is now on par with that of the best PhD students, the report points out that our limited understanding of the implications of AI risks constitutes a "major challenge for policymakers: they must now weigh the benefits and risks of imminent advances in AI without having broad scientific results available". He reinforces the message by saying that the latest trends in AI performance must be an "urgent priority for fundamental research on AI safety in the coming months".

The report goes on to show that pre-emptive risk mitigation based on incomplete or limited evidence may be ineffective or unnecessary. On the other hand, waiting for stronger evidence of the imminent risks that are posed, could leave society caught off guard and even prohibit any further mitigation.

The conclusion of the report could not be clearer: the future of general AIs is uncertain with a wide range of possible trajectories in the near future leading to both very positive and very negative consequences.

These general AIs will be able to be both:

- Engineers: by analyzing a problem, modeling it, optimizing it. They will be able to analyze and define the tasks necessary to achieve the goal.
- Scientists: by inferring models from empirical observations and counterfactual reasoning (events that did not happen but could have been under certain conditions)
- Workers, soldiers, craftsmen: by directing and manipulating physical objects (robots or others) that will themselves use other tools to perform their tasks
- Artists: by creating from scratch new forms of artistic expression in all fields

They will open the field to machines designing and manufacturing other machines, and will create new knowledge, new knowledge, unknown to humans.

Of course, in the midst of these developments there will be fantastic things: new understandings of the universe, of biology, new drugs, less exhausting work, more leisure (as long as we put in place the social safety nets that will be essential and that a real redistribution of wealth is put in place).

But there will also be significant risks. We are already seeing the emergence of emerging self-protection behaviors in current AIs, as Yoshua Bengio says, who claims to have very confidential information on these embarrassing subjects for AI tycoons.

Since we can trust the infinite greed of digital billionaires and the desire for supremacy of autocrats, we can expect that anything that would stand in the way of an unconstrained development of the capabilities of general AIs will be duly combated by all possible means.

AIs are developing very quickly: will they go faster than our ability to supervise them and make us safe in the face of them?

It is therefore absolutely necessary to regulate ex ante, not to wait for problems to arise, to reverse this fatality that has meant that for 50 years, digital regulation has always lagged behind technical developments in the field.

Link to long version of the article :
https://www.academia.edu/128202323/AIs_SAFETY_ISSUES_