# Instinct, language, and artificial intelligence

Gilles Dowek

▶ **To cite this version:**

Gilles Dowek. Instinct, language, and artificial intelligence. 2019. hal-04054594

**HAL Id: hal-04054594**
**https://inria.hal.science/hal-04054594v1**

Preprint submitted on 31 Mar 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Instinct, language, and artificial intelligence

Gilles Dowek[1]

*In the beginning was the Word*

**1. Keeping narratives at a distance.** Our reflection on the epistemology and the ethics of computer science, and of artificial intelligence in particular, is contaminated with a recurring narrative: the creation of human-like living things by human beings. The names of these manufactured doubles proliferate through time: Galatea, the Golem, Frankenstein's creature, Pinocchio, Rossum's robots, HAL 9000… and so does the materials they are made of: ivory, clay, body-parts, wood… But the story remains basically the same with things inevitably going wrong. In the 20th century, robots and computers made their appearances among those creatures without the core of the narrative being affected.

These narratives proceeding from mythology, films, novels, theater plays… raise interesting ethical questions, in particular on the hubris of man defying God by mimicking the Creation and on the boundaries between the human and non-human – a bit like the Valladolid debate in the 16th century. However, they have nothing to do with the epistemological and ethical questions raised by the contemporary developments of artificial intelligence.

As a matter of fact, as soon as names such as Frankenstein or HAL 9000 start looming over this debate, we are not discussing contemporary artificial intelligence at all but this very narrative.

**2. The phrase "Artificial Intelligence".** To designate something that is artificial, it is rather uncommon to associate the adjective "artificial" to a noun referring to something natural, for instance we say "plane" rather than "artificial bird" or "submarine" rather than "artificial fish". The

1 Inria and École normale supérieure de Paris-Saclay.
LSV, 61, avenue du Président Wilson, 94235 Cachan Cedex, France.
gilles.dowek@ens-paris-saclay.fr

May 16th 2019, at the Conference on Robotics, AI, and Humanity, Science, Ethics, and Policy, Pontifical Academy of Sciences and Pontifical Academy of Social Sciences.

association of the adjective “artificial” and the noun “intelligence” for the phrase “artificial intelligence” commonly generates the idea that technological objects are meant to imitate us. This is of course incorrect: cranes, trains, libraries... were not invented to imitate us but to help us to overcoming our deficiencies, for example to lift objects heavier than those we can handle, to travel faster than our legs allow, to record texts that we cannot memorize.

This confusion on the goals of technological objects, the belief that the tools and machines we invent do not complement us but imitate and challenge us, lead generally to a question: “When will the machine surpass us?” If we consider specific tasks – for instance lifting objects, multiplying, playing chess… – tools and machines have already surpassed us for long. Even the fishing-net beats the fisherman at fishing. When trying to extend the problem beyond a specific task, another question arises: “When will the machines become smarter than us?” To formulate this question the introduction of words such as “intelligence”, “consciousness”… is unavoidable. And these words are often used as self-evident whereas, in science and philosophy, they should be questioned in their meaning.

Indeed, when common words or common concepts are introduced in the scientific vocabulary, they are bound to be torn apart: either abandoned – like the concept of phlogiston –, divided into several ones – like the concept of number –, or given a completely different meaning – like the concept of energy.

When the debate on the contemporary development of artificial intelligence starts using the words “intelligence”, “consciousness”… without questioning their meaning we know we have left the realm of science and philosophy.

**3. From “artificial intelligence” to “machine learning”.** The meaning of the phrase “artificial intelligence” has already changed twice in sixty years. Created in 1956, it originally designated the field of computer science that attempted to build programs able to reason logically – later called “automated theorem proving” – in the 1980's this phrase designated the field of computer science that attempted to build programs able to use a natural language – later called “natural language processing”. Today, the phrase designates the field of computer science that attempts to build programs able to learn.

Instead of calling it “artificial intelligence”, we should use the more precise phrase “machine learning”.

**4. Instinct.** We can understand machine learning by describing what it is. But, it is better explained if we describe which problems it solves and which questions it answers. For this, a word which has long disappeared from scientific and philosophical vocabulary will need to be resurrected, the word “instinct”.

For this, let us start an experiment. Picture a little girl knowing nothing about Ambrogio Lorenzentti and Keith Harring and let us show her the data set containing the following pieces of art:

Lorenzetti


Haring


Then show her this other piece

Now, ask her to find out whether it is the work of Lorenzentti or Haring. The little girl identifies Haring's work on the spot. Next, ask her to explain how she managed to do so and she will most probably be unable to describe the process by which she reached this conclusion.

Let us go to this other experiment: ask the same little girl to add 98 and 99, not only will she get the result straight but she will be able to describe the process through which she got to this result: She added 8 and 9 yielding 17, wrote the 7, kept the carry 1…

To use Leibniz's terminology, the first process was “unconscious”: it took place in the little girl's brain without her knowing what was going on. The second one was “conscious” but most importantly it articulated itself with symbols: the digits figuring the numbers. The identification of a particular painting does not use such symbols or if it does, we don't know which ones.

The first process is unconscious and non-linguistic while the second one is conscious and linguistic.

Of course, a neurophysiologist could, in principle, describe in his jargon the state and evolution of each neuron in the little girl's brain when she's identifying the creator of the piece of art, but that language is the one of a neurophysiologist and not the one of the girl. Therefore, although the process can be described externally, in a specific language, this language plays no role in the process itself.

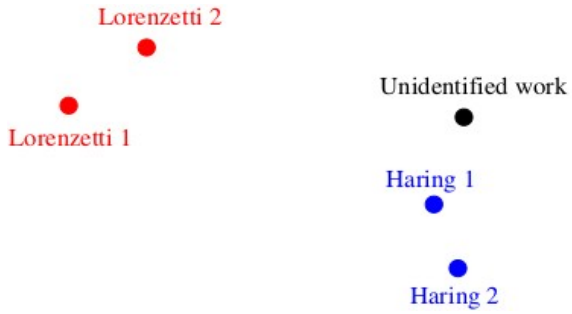
To describe these unconscious non-linguistic processes, now it is time to resurrect the word instinct as we can describe these unconscious non-linguistic process as “instinctive”.

**5. Artificial instinct.** As we can describe how we add the numbers 98 and 99 and how we add two numbers in general, we can program a computer to do it. Still, as we do not know how we identify the creator of a specific piece of art, it is much more difficult to program a computer to do it.

For a long time, such a problem resisted computer scientists: we knew how to build programs that added or multiplied two numbers of one millions digits, which is difficult for humans, but none that distinguished a cat from a dog whereas it is of childlike simplicity.

However solutions were eventually found.

One of these solutions requires defining a metric that measures the similarity between two images which is difficult but not impossible. Then we have to compute the distance of the unidentified work to the four labeled ones of the data set. This shows that it is statistically closer to the works of Haring than to those of Lorenzetti.




So, it is more likely to be a work by Haring than one by Lorenzetti.

Machine learning algorithms can all be seen as an improvement on this very idea, that if something looks statistically like a work of Haring, it is most likely to be a work of Haring. Unlike the algorithm of addition, machine learning algorithms solve problems by using a notion of statistical likelihood without using language.

**6. From instinct to language.** Whenever they perform a computation, computers have an introspective ability to trace it back, step by step, a little bit as if the little girl identifying the author of the painting had the possibility to explain in the language of the neurophysiologist, what happened in her brain. Computers are not black boxes. And processes in computer science are never unconscious.

Therefore, this characteristic can be used to describe in a specific language the process taking place in the machine learning algorithm when identifying Haring as the author of the unknown painting: first it computed the distance of the unidentified work to the first painting of Lorenzetti, then to the second, then it computed the average of these two distances, then…

The process itself is even a little bit more linguistic than it first seemed as computing distances, computing averages, comparing them… uses a language: the language of numbers, even if this language is very simple.

The gap between instinctive processes and linguistic ones suddenly becomes smaller than initially perceived.

**7. Explanation.** Both with the identification of the painter and the addition, the step by step recording of the process is more a description of the process than an explanation of it. In both cases, not much seems to be explained. In the data set the work of art is simply closer to the works of Haring than to those of Lorenzetti and the sum of 98 and 99 is 197, so what is there left to explain?

Long before machine learning, computer science got us used to live with results that had no explanation. For instance, when the weather forecast tells us that, tomorrow, the temperature in Rome will be 25°C, we do not ask for an explanation. We know that this prediction is the result of hours of computation, taking into account thousands of measurements of temperature, atmospheric pressure, wind speed… We could, of course, trace this computation step by step – recording, for instance each step of the solution of a differential equation with the finite element method – but this trace of possibly several terabytes would be more a description than an explanation. And the elements of explanation sometimes given in the weather forecast: “the Azores high got closer to the Iberian Peninsula” are never sufficient to explain such a precise prediction.

There are other cases, however, where even a simple multiplication seems to call for an explanation: for instance when multiplying 12345679 and 36 yields 444444444 – to find the beginning of the explanation, the reader can try to multiply 12345679 and 63.

We can speak of an explanation when we have found a general pattern – such as the fact that multiplying 12345679 and 9*n* where *n* is a digit, yields the number formed with the digit *n* repeated nine times – to which the considered case is an instance.

Of course, such an explanation is difficult to find for a machine learning algorithm. Indeed, recognizing that an unidentified work of art looks more like Haring's works, does not seem to involve any general pattern but only a patient comparison of the image with each image of the data set.

Yet such an explanation can sometimes be found depending on the chosen metric. For instance, if we insist on asking the little girl how she drew a distinction between a piece by Lorenzetti and one by Haring, she may explain that Haring's colors are “brighter” than Lorenzetti's. Now assume that the metric computes three attributes of the images, one of them being its “brightness”, then one can remark that the image to be identified is very bright just like the works of Haring in the data set and that is why it is closer to Haring's work than to Lorenzetti's, the two other attributes playing no substantial role in this evaluation. An explanation could then be “it is more likely to be a work of Haring because the colors are bright”.

But such an explanation is not always possible, in particular with machine learning methods, such as deep learning, with which attributes such as “brightness” are not given by humans but built by the algorithm itself. In this case the attributes of the algorithm don't need to correspond to ours and the expression may be expressed in a language we do not understand – a little bit like with the theory according to which some ancient languages did not have a word for the color blue making the phrase “the blue ocean” impossible to understand for people speaking these languages.

So explaining the result of a machine learning algorithm is still a research problem on which we currently have only partial results. Although there's a lot of progress to come, we must also acknowledge that in some situations, like in the weather forecast, there is nothing to explain.

In such a case, a language is merely a description tool, not an explanatory one.

**8. Why Languages?** If these instinctive algorithms recognize a work of Lorenzetti over one from Haring without manipulating symbols or being able to explain how they reached this conclusion, it becomes unavoidable to wonder: why do we need languages at all?

The first answer could be that in some cases, we do want explanations. For instance, we don't just want to be able to predict the dates of an eclipse, we want to understand why the moon suddenly disappears, which is difficult to explain, and also why the sun disappears, which is easier.

Another reason is that machine learning algorithms rest upon the hypothesis that everything that looks statistically like a dog is most likely to be a dog. There are cases where this continuity assumption is not verified. In these cases, linguistic algorithms must be used. A typical example – that is already in the realm of language – is that everything that statically looks like a correct mathematical proof is not necessarily a correct mathematical proof. Indeed, some incorrect mathematical proofs statically look very much like correct ones. In the same way, everything that statistically looks like a terminating Turing machine is not necessarily a terminating Turing machine. And everything that looks like a bug free program is not necessarily a bug free program.

This seems to lead to the conclusion that there are two branches of computer science, the instinctive algorithms and the linguistic algorithms ones: machine learning and formal methods.

But we should not conclude too hastily as, even if they are fundamentally different, instinct and language can work together.

For instance, if the correctness of a mathematical proof does not rely on instinct, the heuristics used to find such a proof can dramatically benefit from instinct: in an elementary geometry problem, the presence of two orthogonal lines instinctively leads to use the Pythagorean theorem, while the presence of parallel lines instinctively leads to use the intercept theorem, which is statistically a good idea: sometimes a bad one, but often a good one.

**9. From epistemology to ethics.** Moving from epistemology to ethics is also a way to envision a cooperation between instinctive and linguistic algorithms.

Many decisions are taken by algorithms and algorithms often make better decisions than humans. However, can we sentence the accused because his case statistically looks similar to the ones of many convicts? Probably with a strictly consequentialist approach. But not really from a deontological ethics, virtue ethics, or legal point of view. In those cases the judgment must at least be explained by reference to the law article justifying the sentence.

Here the needed explanation is not necessarily the one of the process that leads to the judgment, but the explanation of the reason why the judgment is correct with respect to a legal norm just like a mathematical proof must be correct with respect to a logical norm, independently of the heuristics that led to it.