cooperative approaches

#25 Autumn 2025

Artificial intelligence, collective intelligence

Cooperative approaches

contact@approchescooperatives.com

Cooperative Approaches, a quarterly journal iis published by APAC, a non-profit association based in France. APAC's mission is to promote cooperative approaches in key area of social life. : youth and adult education, social action, organizational management, economy, culture, citizen participation, international life.

English language edition Editor : Larry CHILDS

Editorial board : Matheus BATALHA MOREIRA NERY, Biorn MAYBURY-LEWIS, James ITLO-ADER, David BuLL, Karol QUINN, Dominique BENARD

More information : https://www.approchescooperatives.org/pages/publications-in-english/

Content

CLICK ON A TITLE TO ACCESS THE CORRESPONDING ARTICLE

Artificial intelligence, collective intelligence: forging ahead towards the common good

By Francis Jeandra

We are at a turning point.

A moment when artificial intelligence (AI) is no longer just the stuff of science fiction or pure speculation, but part of our everyday lives: video recommendations, email filters, translation tools, chatbots, optimization logic at every level... AI is everywhere. And yet, few people understand its profound implications, invisible mechanisms, or the real challenges it presents.

It is this intersection between Artificial Intelligence and humanities'Collective Intelligence that this issue of Cooperative Approaches aims to address.

We must not forget that behind every interaction with AI lies very real infrastructure, growing energy consumption, resource extraction, global supply chains, and often precarious workers. In short, a physical world heavily mobilized to produce endless reams of digital material seemingless without effort.

Training models, especially the most powerful ones, requires billions of parameters and hours of computation on energy-intensive servers and data centers, most of which are powered by fossil fuels. It doesn't stop there: their daily, massive use in streaming, automated responses, and predictive analytics perpetuates this dependence on increasingly energy-intensive computing.

So should we throw the baby out with the bathwater? Ban AI? Of course not. But we do need to think differently.

Like any technology, AI is not neutral. It can be a vector for progress or alienation. It all depends on how it is used, how it is governed, and the purposes for which it is assigned.

It can be used to sell more, faster, and at higher prices, often to the detriment of people and the environment. It can also be an ally of algorithmic finance, speculation, targeted advertising, mass surveillance, and automated warfare.

But it can also model the climate with precision, track biodiversity trends, improve access to education, detect weak signals, and be a valuable ally in healthcare, facilitate learning, optimize networks, reduce waste, and more.

Francis JEANDRA

Like any technology, AI is not neutral

There is an urgent need to direct the use of AI towards what makes society work and strengthens our collective ability to face the challenges of our time.

Let's imagine AI that tends less towards endlessly increasing productivity or profit, and rather to making public services more accessible. How about an AI that simplifies procedures, automatically translates into languages with little documentation, identifies situations of vulnerability, or helps explain laws to those who are unfamiliar.

OUR ABILITY TO THINK TOGETHER, TO SET LIMITS AND TO INVENT COOPERATIVE USES WILL MAKE THE DIFFERENCE

Let's imagine a "low-tech" AI that is locally installed, decentralized, operates in a simple and transparent manner, respects personal data, and is understandable and accessible to all.

This requires a real change of course, a shift from the current trend toward massive centralization, proprietary silos, and increased dependence on a few multinational technology companies.

One of the major risks associated with the uncontrolled growth of AI is the increasing automation of decisions and economic exchanges. If poorly regulated, this automation can have powerful and harmful effects on our societies.

On the one hand, digital platforms automate interactions between supply and demand, with algorithms for price setting, inventory management, recommendations, and ratings. This can lead to a dehumanization of commercial relationships, a brutal optimization logic where only the best performers or those ranked highest by the machine survive. The economy becomes rigid, collective deliberation disappears, and the long term ignored.

This is where collective intelligence must come into play.

Faced with increasingly complex systems, it is our ability to think together, to set limits, and to invent cooperative uses that will make the difference. Collective intelligence is the pooling of knowledge, experience, and sensibilities. It is the confrontation of ideas, the development of compromises, and the participation of everyone in defining the tools and rules. It is democracy applied to technology.

Rather than letting AI decide for us, we must decide together what we want to do with it, ensuring its sober and transparent use, so that it becomes a tool for the common good.

AI can be a tool for life—or a cold machine for endless growth for the sake of growth. This choice cannot be left to experts, engineers, or investors alone. It is up to us, collectively, to decide in which direction we want to go.

And that starts now. In our practices, in our demands, in our cooperation.

artificial intelligence is a mirror: it is up to us to choose what it reflects.

Artificial intelligence is entering our lives, our professions, our debates... It fascinates, worries and raises questions. This 22nd edition of Cooperative Approaches (English version) explores what this new form of intelligence can and cannot bring to our collective efforts.

The first part of the issue deals with the fundamentals: How does AI work? What are its promises and opportunities, but also its risks and possible abuses? We address the technical, ethical, and human challenges.

In the second part, we showcase concrete initiatives: Local authorities, researchers, teachers, and how citizens are experimenting with ways to appropriate AI to enrich collective intelligence. From AI cafés to citizens' conventions and from training young people to digital mediation, these approaches show that another relationship with technology is possible, one that instigates more critical thinking, more creative problem solving, and more cooperation.

Happy reading!

BACK TO CONTENTS

FRANCIS JEANDRA

A Scout in France in the 1970s, then a youth leader and educational advisor with SAGIF (Service d'Animation des Golfs d'Ile de France), Francis devoted 14 years to the Scout movement. With a degree in electronics and computer science from EFREI Paris, he had a long career with the Schlumberger group, holding various positions ranging from programmer to work environment director (2009–2021), including training, sales, marketing, quality, health, safety, and the environment.

A recognized expert, he has successfully combined theory and practical experience to effectively support risk prevention in companies.

Elected to the Works Council/Social and Economic Committee/Staff Representative Committee/Health, Safety and Working Conditions Committee between 1992 and 2021, and union representative for the CFE-CGC, he actively contributed to preserving jobs within the group in France during several redundancy plans.

Retired in 2022, he now works as a trainer and consultant in Safety and Environment. Passionate about organic beekeeping, oenology (certified Caviste-Sommelier in 2022) and bonsai, he joined APAC as secretary of the association in 2025.

In our practices, in our demands, in our cooperation, AI is a mirror...

What is artificial intelligence ?

By Dominique Bénard

Artificial intelligence (AI) refers to the ability of a machine, such as a computer, to simulate or reproduce tasks that require human intellectual abilities: performing calculations, analyzing data to make predictions, identifying different types of signs and symbols, conversing with humans, and contributing to the execution of tasks without manual intervention.

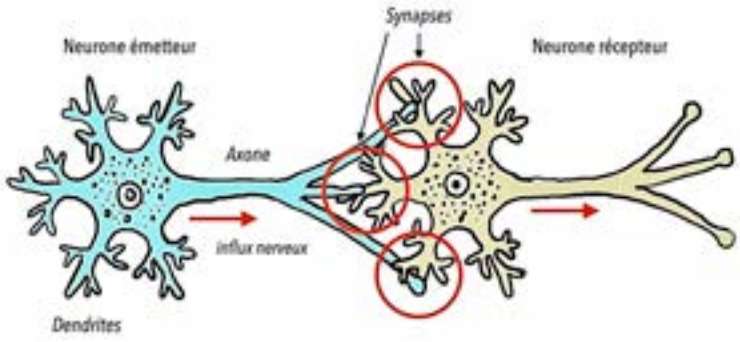
The areas of application for AI are constantly expanding.

HOW DOES AI WORK?

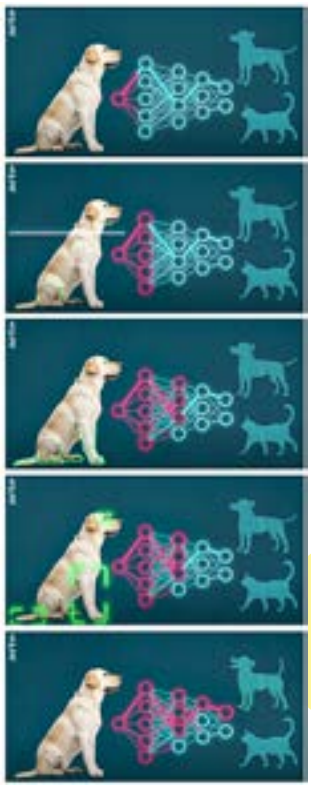
We talk about artificial intelligence because computer engineers were inspired by the human brain, which is made up of nerve cells called neurons that form an extremely complex network. A neuron consists of three parts: the dendrites receive an impulse, which is transmitted via a long cable called the axon, and the synapses connect the neuron to other neurons and enable it to transmit the impulse. This creates circuits that are capable of storing and transmitting information.

In 1943, neurologist Warren McCulloch and logician Walter Pitts examined the question of whether the human nervous system could be considered a kind of universal computing device. They found that when a nerve impulse enters a neuron, if it exceeds a certain threshold, the neuron in turn emits an impulse that is transmitted to other neurons via the axon and synapses. They then made the connection with logic: the impulse or its absence means "activated" or "deactivated," "yes" or "no," 'true' or "false." They realize that a neuron with a threshold low enough that it activates as soon as one of its inputs is activated functions as a physical embodiment of the logical function "or." A neuron with a threshold high enough that it only activates when all of its inputs are activated is a physical embodiment of the logical function "and." They then began to understand that such a "neural network," provided it was properly wired, could do anything that could be done with logic. Together, they wrote a land-

We talk about artificial intelligence because computer engineers were inspired by the human brain...



Synapses enable neurons to connect to form networks in which nerve impulses circulate.



mark scientific article entitled, "*A Logical Calculus of Ideas Immanent in Nervous Activity*." This article is the basis for artificial neural networks.

Computer engineers did not try to copy the biological network because it is too complex. They programmed an abstraction of how it works to build *artificial neural networks*. These networks are not comprised of neurons but of several computer programs (or algorithms) running in parallel.

ARTIFICIAL NEURAL NETWORKS

An artificial neural network is not a physical circuit of cables and processors, but a set of computer programs simulating neurons arranged in layers (hence the term "deep learning") and connected to each other in a chain. If the analysis suggests a dog, this activates other neurons that will search for even more complex canine characteristics. This chain reaction ultimately produces a result, or rather a prediction: it is probably a dog (see diagram above). The brain of an AI therefore exhibits a pattern of activity that varies depending on the task it is performing at a given moment.

TRANSFORMERS AND SELF-ATTENTION

To be able to interpret language, LLMs need to understand more than just isolated words. They need to be able to interpret sentences, paragraphs, and entire documents. Early machine learning models struggled to understand entire sentences and generally "forgot" the beginning of a sentence by the time they reached the end, leading to misinterpretations.

Modern generative AI models use a specific type of neural network called transformers. These perform a feature called self-attention to detect how elements in a sequence are connected. Transformers enable generative AI models to process and contextualize large blocks of text rather than isolated words and sentences.

Example: we want to train a model to recognize grammatical patterns such as "a cat sleeps." In a traditional artificial neural network, each layer processes a different part of the information: words, their relationships, their order, and so on. This is called hierarchical learning. Before 2017, recurrent neural networks (RNNs) were mainly used. These models analyze sentences word by word, from left to right, which limits their ability to understand complex relationships between words that are far apart in a sentence. For example, in the sentence "The dog, which is sleeping, is sleeping," the word 'which' is not recognized because it is not part of the sentence "The dog is sleeping." (RNNs). These models analyze sentences word by word, from left to right, which limits their ability to understand complex relationships between words that are far apart in a sentence. For example, in the sentence "The dog that Marie adopted yesterday is sleeping on the sofa," RNNs had difficulty connecting 'dog' to "sleeping."

This is where the transformer comes in to change the game.

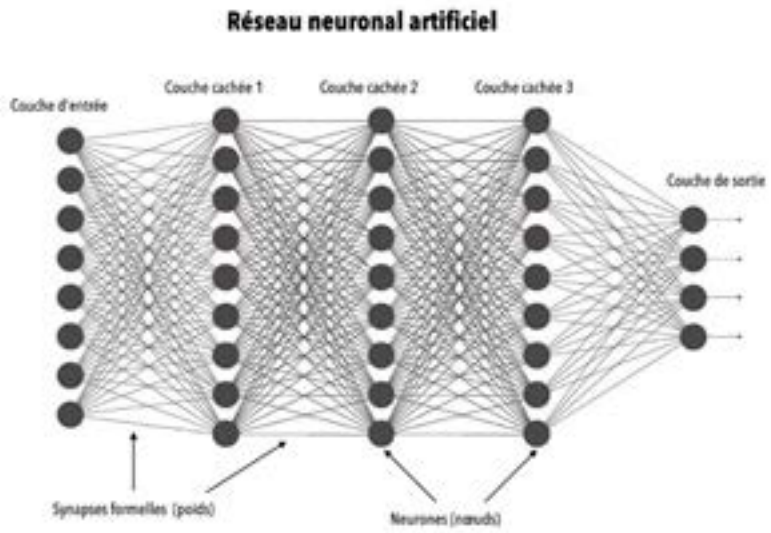
Example: "*Marie loves her brother and supports him in everything he*

does." When we read this sentence, we know that "*she*" refers to "Marie" and 'him' refers to "brother." But for a computer, this kind of connection is not obvious. The self-attention mechanism allows the model to look at each word and calculate its importance relative to the other words in the sentence. In other words, it assigns a weight (a score) to each word to decide which elements of the text are most relevant. In the sentence "Marie loves her brother and she supports him in everything he does," the word "she" will have a strong link to "Marie," while 'he' will be linked to "brother." Thanks to this ability, transformers can analyze an entire sentence at once instead of processing it word by word. This allows them to better understand complex relationships in a text. This mechanism drastically improves the consistency and accuracy of models. It allows long sentences or entire texts to be processed without losing the thread of the context.

LLM TRAINING

For a model to become effective, it must go through two training stages: pre-training and fine-tuning.

Pre-training: the model is exposed to a huge amount of text from books, articles, web pages, etc. The goal is for it to learn patterns, i.e., recurring patterns in the data, which are determined by the fundamental rules of language: grammar, syntax, and frequent associations between words. If the model often sees the sentence "the cat is sleeping on the sofa," it learns that the words "cat," "sleeping," and "sofa" are often associated. It does not understand what a cat or a sofa really is. To give an idea of the scale of this learning process, some models are trained on trillions of



words. It's like reading the equivalent of 16 million books.

Refining: in this second phase, the model is trained on specific data for specific tasks. For example, a model used in the medical field will be refined with scientific articles and clinical studies. Refinement allows the model to be adapted to specific contexts. If, before refinement, we ask the question "What is the flu?", the model might respond: "a common illness caused by a virus." After refinement, it might respond, "The flu is caused by the influenza A or B virus and causes fever, muscle aches, and chills." This level of precision is possible thanks to targeted training data that allows the model to become an expert in a particular field.

With transformers, much larger models can be created that are capable of learning from massive amounts of data. For example, OpenIA's GPT3 uses 175 billion parameters. By comparison, GPT2 had "only" 1.5 billion, and GPT4 is expected to use 1 trillion. This increase in size allows for the capture of more complex nuances in language. Thanks to their versatility, transformers are not

FOR A MODEL TO BECOME EFFECTIVE, IT MUST GO THROUGH TWO TRAINING STAGES...

limited to text comprehension; they can also be used for other tasks such as image recognition, with variants such as vision transformers, or music generation.

Training LLMs consumes large amounts of energy and results in massive CO2 emissions.

THE VOLUME OF TEXTUAL INFORMATION THAT A LLM DIGEST TODAY IS EQUIVALENT TO WHAT A HUMAN CHILD RECEIVES SOLELY THROUGH THEIR EYES IN THE FIRST FOUR YEARS OF LIFE...

ARE LLMs AT A DEAD END?

The current strategy of AI companies is to build ever larger LLMs, working on ever more data, with ever more computing power, in the hope of bringing about the famous AGI, or artificial general intelligence, which will surpass humans in every way.

According to renowned comp ter scientist Yann Le Cun[1], this is a monumental mistake. Training human intelligence from text, as LLMs do, will never be enough to achieve human intelligence. Yann Le Cun points out that the volume of textual information that a large language model (LLM) digests today is equivalent to what a human child receives solely through their eyes in the first four years of life. He emphasizes that language alone is not the whole of intelligence and argues that it would be more interesting to look at how to enable machines to develop a model of the world in order to understand the physical world, so that they are capable of reasoning and planning, rather than simply calculating the probability of the next words appearing in a text.

He explains: «*A model of the world is what we all have in our minds that allows us to manipulate thoughts. I know that if I push the top of this bottle, it will probably tip over, whereas if I push the bottom, it will slide across the table. We have models of the physical world that we acquire in the first months of our lives, and that's what allows us to deal with the real world. It's much harder than dealing with language. True AI requires models that can deal with the real world*.»

He proposes creating systems that would be capable, like us, of understanding the physical consequences of actions, not just capable of manipulating language.

Yann Le Cun's point of view is corroborated by Judea Pearl, a renowned computer scientist and philosopher, and a leading expert in causal reasoning, who wrote in his book, The Book of Why: Systems capable of intelligent action must be able to manipulate causal representations of the world, not just correlate observations.

1. In the 1980s, Yann Le Cun, a French computer engineer, began working on machine learning. He wanted to teach machines to recognize images, sounds, and even text by showing them thousands of examples. He developed the key concept of the convolutional neural network, which is still the basic technology behind AI today. In 2018, he received the Turing Award, the equivalent of the Nobel Prize in computer science, alongside Geoffrey Hinton and Yoshua Benjo. He is the chief AI scientist at Meta, one of the most strategic positions in the world of AI.

THE MYTH AND REALITIES OF ARTIFICIAL INTELLIGENCE

Another famous computer scientist, Luc Julia[2], denounces the myths surrounding AI. According to him, far too many people talk about artificial intelligence without knowing what they are talking about. They spread a "Hollywood" vision of AI: something quite fantastical, frightening or dreamlike, an artificial entity that will surpass human intelligence, take our jobs, and kill us (The Matrix, Terminator, etc.). Luc Julia states: "*There is no single artificial intelligence that can do everything; there are several. They are very specific tools that do very specialized things. AI does things better than you, just like any other tool. A hammer drives nails better than you. But as with a hammer, with AI, you're the one holding the handle, you're the one making the decisions.*"

In Le Monde on December 30, 2024, Daron Acemoglu[3], professor of economics at the Massachusetts Institute of Technology (MIT), denounced the dream of general artificial intelligence capable of performing any human cognitive task. According to him, the obsession with building super-intelligent machines is leading the industry to

ignore the real potential of AI. Instead of pursuing such a quest, it would be more reasonable to develop a different kind of AI that is technically feasible and socially desirable. Instead of seeking to replace workers, it would be wiser to use AI to help them perform difficult tasks.

"We need an industrial paradigm," he writes, "that, rather than celebrating the superiority of machines, emphasizes their main advantage: the improvement and expansion of human capabilities."

To achieve this, it is important to invest more in training and developing human skills so that workers are able to understand, use, and master AI systems.

THERE IS NO SINGLE ARTIFICIAL INTELLIGENCE THAT CAN DO EVERITHING... THERE AURE VERY SPECIFIC TOOLS THAT DO VERY SPECIALIZED THINGS...

BACK TO CONTENTS

2. Luc Julia: born in January 1966 in Toulouse, is a French-American engineer and computer scientist, researcher (CNRS, Stanford) specializing in artificial intelligence. He is one of the designers of the Siri voice assistant. Former vice president of Samsung in charge of innovation, he joined Renault in 2021 as chief scientific officer.

3. Daron Acemoglu, winner of the 2024 Nobel Prize in Economics, is the author, along with Simon Johnson, of "Power and Progress: Our 1000-Year Struggle Over Technology & Prosperity."


1010110110101010110
SCANNING

ARTIFICIAL INTELLIGENCE: A NEW COLLABORATOR IN COLLECTIVE INTELLIGENCE?

By Laurent Butré

Since humanity discovered fire and worked stones to make tools, it has never stopped inventing. Each generation has seen technological revolutions that have transformed the way we live, think, and work.

A REVOLUTION EQUIVALENT TO THAT OF THE INTERNET IN THE 2000'S

For thousands of years, our ancestors progressed slowly: a better-cut flint here, a better-fixed wheel there... Then, suddenly, in the 18th century, everything accelerated with the Industrial Revolution. Steam engines, spinning mills, locomotives: humans began to delegate their muscle power to machines. The world changed. And this change accelerated even further with electricity, the telephone, and then computers.

But another technological earthquake struck at the end of the 20th century: the arrival of the internet. In the 2000s, the web became accessible to everyone. We can chat with someone on the other side of the world, buy a book without leaving our sofa, and find information in just a few clicks. The world has become a connected village. It's a radical transformation. Some call this period the "third industrial revolution."

And now? We are experiencing a new turning point, another disruption: the arrival of artificial intelligence (AI), more specifically generative AI. This is our era, our upheaval. The tool is changing shape once again: it is no longer just a machine that obeys orders, it is an entity capable of proposing, understanding, creating... just like a human being.

For the first time, machines are beginning to think with us. And I don't really like the term "think" because behind it lies the understanding that it is only statistics based on what has already been created.

COLLABORATE WITH IT, COAX IT, LEARN TO USE IT

If we look around us, artificial intelligence is already everywhere. It's not something new for 2023 or 2024. For several years now, our phones, cars, and computers have been using very powerful algorithms, some of which are invisible.

Laurent BUTRE

Do you use GPS? It calculates the fastest route in seconds. Before, you had to unfold a map, find your location, and guess.

Want to know what song is playing in a café? Shazam listens to it and recognizes it.

ARTIFICIAL INTELLIGENCE, DESPITE ITS NAME, IS NOT CONSCIOUSNESS. IT IS AN EXTREMELY POWERFUL STATISTICAL TOOL...

Your phone corrects your spelling mistakes, anticipates what you want to write, and suggests images from your gallery by recognizing faces or places.

All of this is already artificial intelligence. Discreet. Practical. But what's changing today is that this intelligence is becoming more accessible, more widespread, and more creative. We're no longer talking about a tool that does one specific thing.

AI can now write a text, create an image, suggest a recipe, analyze a contract, generate computer code, summarize an article, plan an event... And we have to learn to work with it.

Gartner, a leading American consulting firm, has made it clear:

«*AI will not take people's jobs, but those who do not work with AI will no longer have jobs.*»

As with any innovation, those who know how to use it get ahead. This is true for students, teachers, researchers, writers, lawyers, doctors, engineers... AI is becoming a work companion. And to collaborate effectively, we need to understand it, tame it, and use it intelligently.

WHAT IS ARTIFICIAL INTELLIGENCE?

So what exactly is this famous AI? Is it a conscious machine? Does it really think?

No. Artificial intelligence, despite its name, is not consciousness. It is an extremely powerful statistical tool.

It can be defined as follows:

"*It is the ability of a computer program to observe its environment, learn from it, reason about what it perceives, decide as a human would, and, above all, act.*"

But what does that mean in practice? Let's take an example. You show an image to an AI so that it can analyze and use it, and it transforms it into numbers. The same goes for sound and text. Everything becomes mathematical and measurable. It then calculates probabilities: based on millions of past examples, what are the chances that this signal corresponds to a cat, a musical note, or a spelling mistake?

For this to work, you need a huge amount of data. AI doesn't "understand" anything: it learns from examples. The more examples you give it, the better its predictions will be. So it's not intelligence in the human sense, It's intelligence through statistical imitation. And everything it does boils down to solving two types of problems:

• *Classification problems*: For example, recognizing whether an email is spam or not. It's like choosing which box to put a piece of data in.

• *Regression problems*: For example, predicting the price of a house based on its size, location, etc. We look for the most likely value within a range.

And the best part is that any question can be rephrased to fit into one of these two categories. This is the power of machine learning, and especially of data scientists, who are data engineers who spend their time thinking about how to rephrase problems using these two approaches!

WHAT IS GENERATIVE ARTIFICIAL INTELLIGENCE?

And here's the surprise: we learn that these machines can generate text, create images, and compose music! How is this possible, if all they do is classify or estimate probabilities? This is where the strength of researchers comes into play.

Let's take the example of a text. We ask the AI a seemingly simple question:

"*Complete this sentence: Artificial intelligence is...*"

But in reality, we rephrase the problem as follows:

"*Among a million possible words, which one is most likely to be the next word that completes the three dots in the sentence above?*"

And there you have it! It's classification.

The AI chooses the most likely word. Then it starts again, with the new word added to the sentence. Word by word, it "generates" a text. The same goes for an image: pixel by pixel, area by area, it predicts what is most likely to be there.

But to achieve this result, it needs to have read, seen, and heard millions, even billions, of pieces of content. This is what OpenAI, Google, Microsoft, Mistral, and others have done: they have sucked up the web, books, articles, forums, databases...

Problem: depending on the sources used, the answers will not be the same. This is called model bias. An AI that has been trained mainly on American texts will have a different "culture" from an AI trained on French-language, scientific, or artistic data. Hence the importance of picking and choosing from the various offerings, including ChatGPT, Gemini, Mistral, Bing CoPilot, Claude, Llama (Meta), Grok, and others.

IT'S HERE TO STAY... AND GO EVEN FURTHER

Some criticize AI for its environmental impact. And it's true: training a large AI model requires a lot of energy. Gigantic data centers, servers that heat up, billions of calculations...

But today, we can no longer turn it off. Why? Because it makes those who use it more efficient. It saves time, increases productivity, and stimulates creativity.

AI can make mistakes, invent things, and hallucinate. And sometimes

THE AI CHOOSES THE MOST LIKELY WORD. THEN IT STARTS AGAIN, WITH THE NEW WORD ADDED TO THE SENTENCE. WORD BY WORD, IT «GENERATES» A TEXT...

© Shutterstock / Stock Asso - Agent IA

it does so with such aplomb and confidence that we believe it! This is dangerous.

Why? Because it all depends on the data it has seen. And since it doesn't understand, it can mix up true and false information. Worse, it can fabricate quotes, invent facts, and assemble incoherent pieces of sentences.

THE NEXT STEP IS THE ARRIVAL OF INTELLIGENT AGENTS. UNTIL NOW, GENERATIVE AI HAS BEEN ABLE OF CREATING CONTENT. BUT NOW, IT WILL TAKE ACTION...

It's like a GPS. It can guide you down a one-way street. But you're not going to listen to it with your eyes closed. You keep your critical thinking skills sharp and watch the road. With AI, it's the same: you check, cross-check, and compare. AI is an assistant, not an oracle.

AGENTS... AND AGENTS OF AGENTS

That's why we need to learn how to use it properly, especially by mastering the art of prompt engineering.

Before, with Google, we would type in three keywords. And the search engine would find the sites that contained the information we were looking for because they were better referenced, linked, viewed, etc. Now, we no longer search, we talk to AI. And since it completes the text, the more precise, complete, and context-rich your question is, the more relevant the answer will be.

An effective prompt is a well-worded prompt. Example:

- ***Bad prompt***: *"Write an article about AI."*
- ***Good prompt***: *"Write a 1,000-word article on the impact of AI in healthcare, using an educational tone, aimed at high school students."*

The clearer you are, the more surprising and sometimes impressive AI can be.

MAINTAIN YOUR CRITICAL THINKING

But be careful. These are only statistics. They are not absolute truths.

The next step is the arrival of intelligent agents. Until now, generative AI has been capable of creating content. But now, it will take action.

An agent is a program capable of using AI responses to perform a task, interact with other services, and make decisions. And if we combine them, we can get them to solve complex problems.

Imagine you want to organize a vacation:

1. A first agent talks to you to understand what you want: beach or museum? France or Japan?
2. Another agent optimizes the itinerary according to your tastes.
3. A third agent books the train or plane tickets.
4. A fourth looks for the best hotels.
5. Another manages your schedule and adds the addresses to your GPS...

And all this without you having to do anything, except express your needs at the beginning.

This is the future of agents. Specialized intelligences that collaborate with each other to respond to your requests. Like a virtual team at your service.

WILL AGENTS BECOME A COLLECTIVE INTELLIGENCE?

We're almost there. Artificial intelligence is here. Not to replace us, but to support us. Well, we hope so, anyway. We've all seen the disaster scenarios in movies over the last 30 years. But artificial intelligence is like a new colleague. A powerful assistant. A creative partner. A fast executor.

But beware: it doesn't think for us. It has no conscience, no values, no intuition. What makes us human is our imagination, our critical thinking, our ethics. We must also be mindful of the bias that every AI-creating company can introduce into its algorithms to influence, control, and guide us, whether intentionally or not.

AI will become part of humanity's collective intelligence. It will be one voice among many.

But it is up to us, as humans, to remain in control. To ask the right questions. To dream. To invent. To doubt.

This new turning point is irreversible. So we might as well learn to embrace it together, with clarity, curiosity... and a healthy dose of critical thinking!

BACK TO CONTENTS

Laurent Butré is a seasoned digital expert with nearly three decades of experience at SLB (formerly Schlumberger).

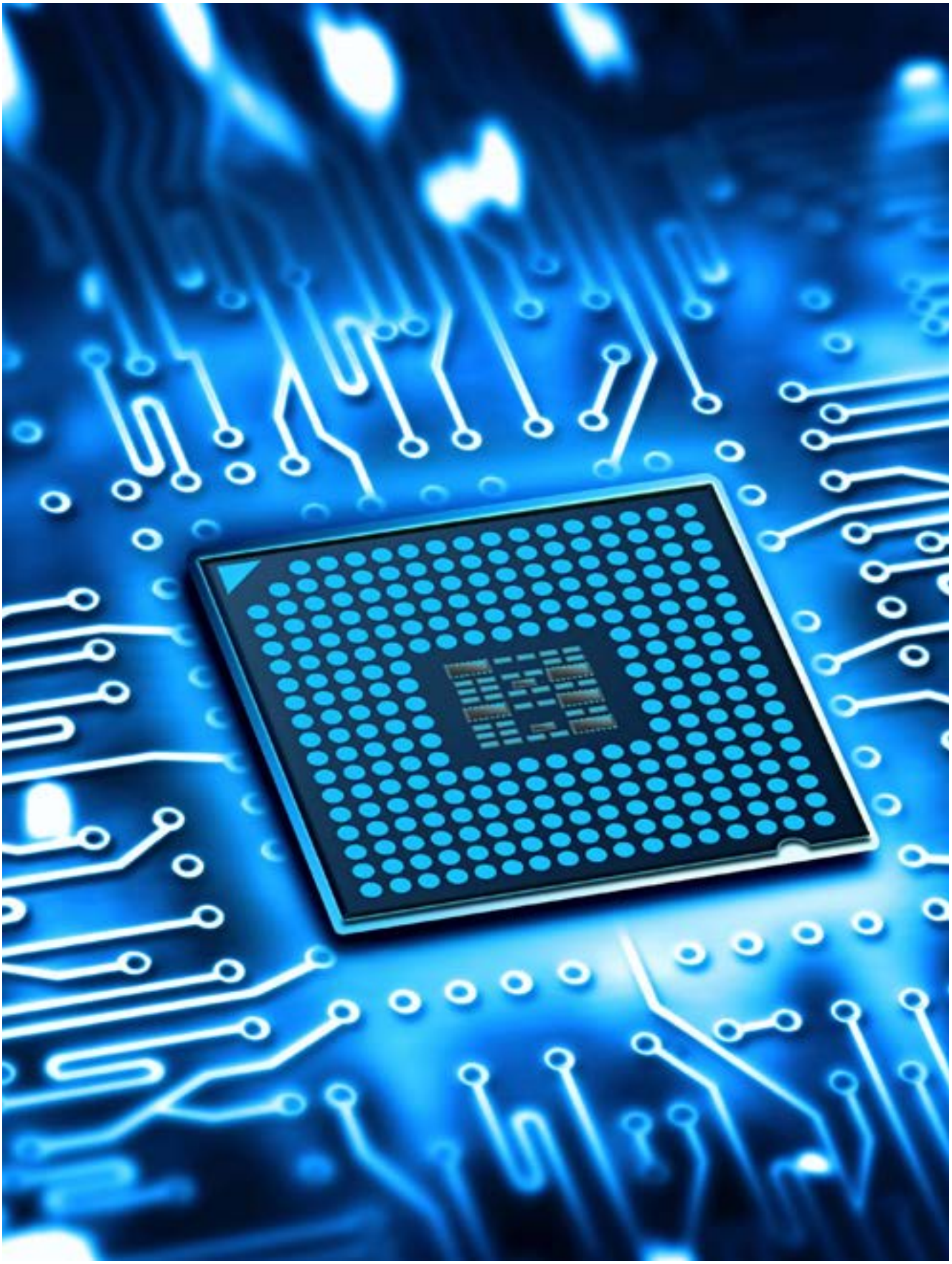
From his early days in IT support to leading global digital operations as Digital Operations Director, he has spent more than 15 years working internationally.

Having held key roles in Australia, Texas, and Silicon Valley, he is currently based in Paris after returning to France to start an Artificial Intelligence Lab in Paris (Clamart, 92).

Laurent has always had a guiding principle: simplify technology to maximize impact: "Make it simple; initiate purposeful impacts."

A graduate of ENSIMAG, he also sits on the board of directors of IHES.

What motivates him? Creating solutions that save time and make a real difference to users' lives.



FIVE MAJOR CHALLENGES FACING AI

By Sylvestre Bénard

In 2024, Geoffrey Hinton and John Hopfield jointly received the Nobel Prize in Physics for their work on deep neural networks at Google's laboratories. They are considered by the community to be among the fathers of generative AI.

In addition to the fact that many physicists were surprised that the prize was awarded to computer scientists, the notable fact about this edition is that Geoffrey Hinton had resigned from Google a few months earlier in order to be able to freely warn the public about the dangers of generative AI. At the award ceremony, he spoke at length about the problems posed by AI: the replacement of technical and creative jobs, criminal or illegal use, and the existential risk posed by general artificial intelligence.

It is not uncommon for renowned scientists to warn of the dangers of technologies they have helped to develop. We naturally think of Einstein, who spoke at length about the major risk that the atomic bomb posed to humanity.

Artificial intelligence is now ubiquitous, used by institutions, politicians, the armed forces, advertisers, journalists, and artists alike. Is it really risk-free?

In this article, I invite you to take a step back and look at five challenges that generative artificial intelligence will have to overcome if it is to become a mature technology that is useful to everyone: the challenges of energy, data collection, fairness and bias, hallucinations, and finally, the social context.

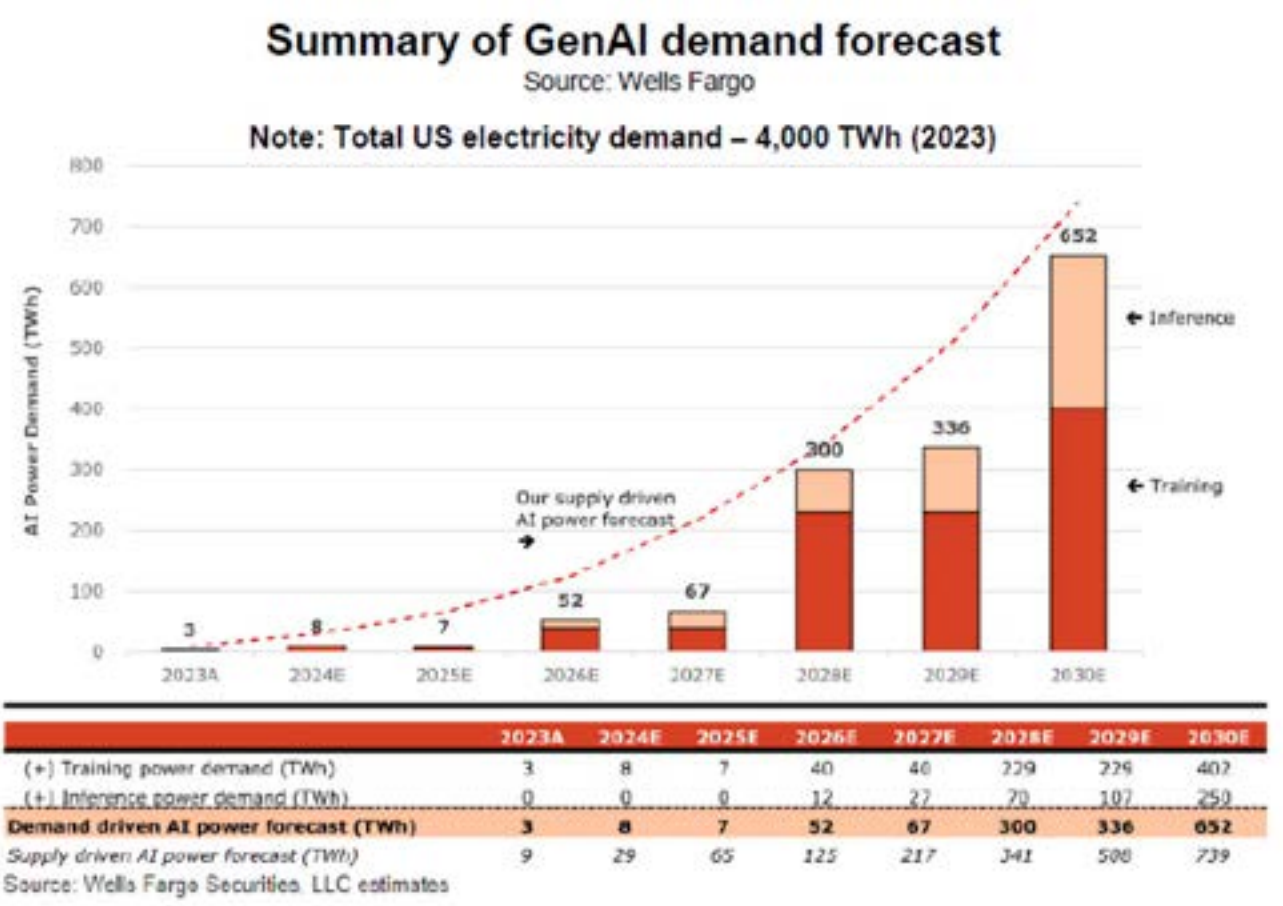
Sylvestre BENARD

THE ENERGY CHALLENGE

Just as Geoffrey Hinton received his Nobel Prize, world leaders gathered in the opulent hotels of Davos. For several years, artificial intelligence had been an important topic of discussion at the Forum. On this occasion, Sam Altman, CEO of OpenAI—the developers of the famous ChatGPT—gave a lecture on the future of artificial intelligence. Regarding the energy requirements for the development of general AI, he stated:

«*There is no way to achieve this without a breakthrough. This motivates us to invest more in nuclear fusion.*»

He had thrown a cat among the pigeons, and journalists realized that energy was going to be one of the major challenges for the growth of AI applications.



	2023A	2024E	2025E	2026E	2027E	2028E	2029E	2030E
(+) Training power demand (TWh)	3	8	7	40	40	229	229	402
(+) Inference power demand (TWh)	0	0	0	12	27	70	107	250
Demand driven AI power forecast (TWh)	**3**	**8**	**7**	**52**	**67**	**300**	**336**	**652**
Supply driven AI power forecast (TWh)	*9*	*29*	*65*	*125*	*217*	*341*	*500*	*739*

Source: Wells Fargo Securities, LLC estimates

Artificial intelligence is forecast to require much more energy in the years ahead.
PHOTO: AEP, WELLS FARGO

It is important to note that a query on a generative AI such as Grok, ChatGpt or Google Gemini consumes on average 10 times more energy than a simple Internet search. But that's just the tip of the iceberg. The energy consumption of generative AI can be divided into two phases:

- ***the training phase***, during which it ingests enormous amounts of data to build its model,
- the ***operational phase,*** during which it uses its neural network to respond to queries.

The training phase requires enormous computing power. For example, Grok 2, Elon Musk's AI, required 20,000 Nvidia H100 computing processors, and Grok 3 required nearly 100,000. Each H100 processor consumes between 350 and 700 watts. In total, this corresponds to an instantaneous consumption of 35 to 70 megawatts, which is equivalent to the consumption of a city of 4,000 inhabitants. This phase accounts for nearly two-thirds of the total energy consumption during the lifetime of an AI. But Sam Altman was not taken seriously. Indeed, with nuclear fusion still a distant dream, the CEO of OpenAI was considered a provocateur.

However, in June of the same year, Microsoft signed an agreement to restart the second reactor at the infamous Three Mile Island nuclear power plant, which had been shut down since 2019 (the first reactor was the site of the worst nuclear accident in US history in 1979). The company then stated that the energy produced (850 MW) would be used for its data centers (whose growth is closely linked to the development of their AI services). Two years later, buoyed by Donald Trump's "Stargate" mega-project, FERMI launched Hypergrid, a huge AI-dedicated data center in Texas that will generate its own electricity (up to 11 GW) by combining nuclear, solar, and gas power.

A projection by Wells Fargo shows the scale of this: in 2030, all AI-related activities worldwide could represent 652 terawatts, more than France's total annual production in 2023 (519 TW).

We often hear AI advocates say that humans whose jobs are replaced by AI also consume energy. So the imbalance would not be that significant. However, if we go back to 2015 when Google's AI beat world Go champion Lee Sedol, Google's "AlphaGo" computer consumed 15,000 W, while Lee Sedol's brain consumed a maximum of 50 W (20% of the energy consumption of an adult male).

THE CHALLENGE OF DATA COLLECTION

Generative AI is not only an energy guzzler, it is also a data hog. While a human being can reproduce a task with a few examples and draw logical rules from past experience, generative AI must be trained using millions or even billions of contextualized data points.

As a result, the lack of sufficient training data has long been a barrier to the development of deep neural networks. The turning point came shortly after 2006 with the rise of Web 2.0. Social networks built up mountains of user data and already had a business model centered on exploiting it.

It was thanks to researcher Fei-Fei Li at Stanford University that a shift took place in 2010. As a computer science student, she realized that the lack of high-quality data was hindering the development of image recognition by neural networks. She then used Amazon's "Mechanical Turk" (a network of workers in poor English-speaking countries paid by the task) to classify up to 16 million images. Between 2010 and 2016, she launched an image recognition competition that helped refine modern neural network models.

In the following decade, AI-based generative applications flourished, requiring ever more data. DeepL, for example, used all the transcripts of the European Parliament and all the books translated and available online.

Large generative models such as Chat GPT and Midjourney will begin scanning the entire web to feed their increasingly data-hungry training. The latest model from Dall-E (one of the leaders in image generation) is based on nearly 5.8 billion annotated images, and the amount of data required to train ChatGPT has increased tenfold in four years.

In 2030, all AI-related activities worldwide could represent 652 terawatts, more than France's total annual production...

In this context, another problem is beginning to emerge: the scarcity of high-quality data. This is impacting all AI players. The need for data is such that companies developing AI models are capturing all available data regardless of its source and without concern for image or copyright issues.

They are starting to use AI-generated content for their own training.

In 2024, according to a study by Amazon Web Services, nearly 57% of data posted online will have been produced by AI. On the social network LinkedIn, the agency Influence Metrics has shown that 61% of posts come from generative AI. This creates a vicious circle that amplifies learning defects.

Alongside this impoverishment of data, access to public data as become more restricted...

Alongside this impoverishment of data, access to public data has become more restricted, with many sites deciding to prevent AI from plundering their content. All over the world, rights holders are organizing to resist this plundering.

Legislation is becoming more protective. In 2023, the Biden directive aims to protect citizens from the excesses of AI. In Europe, the GDPR and the Artificial Intelligence Act impose even more restrictions on AI players.

With this restrictive environment, some analysts predict a slowdown in development and a decline in the most resource-intensive LLMs[1].

However, resistance from the AI giants is mounting. In 2024, META and around 30 other companies sent an open letter to the European Commission asking it to review the regulations.

Several experts also believe that Elon Musk's real goal with the Department of Government Efficacity (DOGE) was to get hold of data from US government departments (high-quality, perfectly categorized data) to feed his company xAI and its chatbot Grok[2].

THE CHALLENGE OF BIAS AND EQUITY

In 2013, Google, then at the forefront of deep neural network technology, launched Word2Vec. This program classified words as vectors in a 300-dimensional space. It then became possible to perform Boolean operations on words, with the crossing of vectors revealing the relationships between them.

For example:

- Japan: Tokyo :: France: Paris
- man: king :: woman: queen

Google's article had a huge impact because the "word embedding" technique was a test used by neuropsychologists to assess the creativity of their patients.

For the first time, computers seemed to be able to match a higher form of human behavior.

However, one afternoon, researcher Tolga Bolukbasi and a colleague were killing time in the cafeteria at Boston University. They were using Word-

1. https://www.nytimes.com/2024/07/19/technology/ ai-data-restrictions.html

2.. https://theconversation.com/doge-threat-how-government-data-would-give-an-ai-companyextraordinary-power-250907

2Vec on their smartphones and discussing the technology. Impressed by the power of the system, they tried different word combinations. By chance, he typed in the following query without thinking too much about it: father - doctor; mother... and the system returned nurse.

Half surprised by the response, he tried two more queries

- building - man: architect
- building - woman: cleaning lady

Tolga Bolukbasi had just discovered that Google's system was sexist[3].

It turned out that the corpus on which Google had trained its system contained misogynistic texts, which the system had integrated without the programmers' knowledge. This was one of the first examples of how artificial intelligence systems based on neural networks could be biased and even seem to accentuate bias.

While this discovery concerned academic research, it didn't take long for the technology's shortcomings to become apparent in society. In 2014, Amazon's recruitment team in the US decided to secretly implement an AI system to sort candidates' resumes. They trained the system using their resume database and launched it in a test phase. The system seemed efficient and recruited competent people.

However, after a few weeks, management realized that they were no longer hiring any women. They analyzed the neural network and saw that it created a strong link between male first names and competence. They therefore decided to disregard first names. But the AI continued to behave in the same way. They analyzed the network again and discovered that women did not choose the same specialties in their training. Similarly, women did not play the same sports or have the same hobbies.

The system had amplified the biases and encoded them through multiple links that were so intertwined that it was impossible to correct. Amazon had to shut down the project[4].

Many other examples of bias caused scandal, such as people of color being classified as gorillas by Google Images, or facial recognition systems that performed between 40 and 100% worse for black women than for white men, as demonstrated by Dr. Joy Buolamwini, founder of the Algorithmic Justice League, in her research.

Deep neural network systems are black boxes that create invisible correlations between concepts and provide answers that amplify majority connections, thereby amplifying sexist and racist biases[5].

This "black box" syndrome also makes it impossible to track and control these connections in large models such as ChatGPT, as they contain hundreds of billions of them.

IT TURNED OUT THAT THE CORPUS ON WICH GOOGLE HAD TRAINED ITS SYSTEM CONTAINED MISOGYNISTIC TEXTS...

3. https://www.technologyreview.com/2016/07/27/158634/how-vector-spacemathematics-reveals-the-hiddensexism-in-language/

4. https://www.reuters.com/article/world/insighta-mazon-scraps-secret-ai-recruiting-tool-thatshowed-bias-against-women-idUSKCN1MK0AG/

5. https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms

Image generators still have a hard time generating realistic hands...

Today, ask Midjourney, for example, to create an architect's office, and it will display a male architect 100% of the time. Yet a simple Google image search will easily show you 40% female architects. This is a perfect example of bias amplification.

But biases are not just a training error or a flaw in the technology. They perpetuate power relations. If we take a step back, the engineers who developed these AIs are mostly young white men, and they have created tools that reflect their own image.

This tendency to favor majority data also poses a major problem for the representation of minority cultures that do not have enough written data to train a translation AI or LLM.

THE CHALLENGE OF HALLUCINATIONS

The other major problem with generative AI, which is just as difficult to get around, is its tendency to hallucinate or "co-fabulate" when it offers a false or misleading answer. We can all experience this by asking specific questions about data that the chatbot does not have. It will never say that it cannot answer, but will instead generate perfectly plausible but false content. It will always display an excess of confidence.

In 2023, Teresa Kubacka, a doctor of data science, asked ChatGPT to give her the definition of "cycloidal reverse electromagnet," a term invented for the occasion. The chatbot gave a perfectly plausible definition, but it contained false studies, false dates, and false authors. The false citations were so credible that the researcher had to check them to be sure.

Similarly, image generators still have a hard time generating realistic hands. This is because no one photographs their hands in different positions and adds comments.

There are therefore no photos of this type on Instagram, for example.

This well-documented problem discredits generative AI tools, and companies are trying to combat this behavior. Some use pairs of antagonistic AI. This is the case with OpenAI's ChatGPTpro, which costs $200 per month. Both AI systems give an answer, and if they are too different, it is considered a hallucination.

Others have discovered that disabling parts of the neural network during training limits the cross-links that create correlations between very distant concepts. However, no solution has yet been found to resolve this issue. It appears to be inherent to the way generative AI works.

THE CHALLENGE OF SOCIAL ACCEPTANCE

As we have seen, artificial intelligence is a technology that is far from mature. Yet it is increasingly being used by many players, sometimes

for critical tasks that can have serious consequences for citizens' lives.

In 2023, one of the largest strikes in the history of cinema and media began in the US: the writers' strike. It was largely presented as a wage dispute, but one of the main reasons for the mobilization was the widespread use of chatbots by studios. For the first time in history, machines were threatening high-level creative jobs.

In fact, it was during this period that companies began to automate back-office tasks with AI. A publicity campaign in the streets of San Francisco caused quite a stir: posters from the company Artisan called for companies to stop hiring humans and replace them with their AI agents.

Since then, the trend has only grown stronger. Amazon began reducing its workforce in 2025 as its revenue skyrocketed. According to Jeff Bezos, "*thanks to AI, we will need fewer people for the jobs we do today*."

Luis von Ahn, CEO of Duolingo, decided that bonuses would only be distributed to teams that automate some of their tasks. Similarly, Tobi Lütke, CEO of Shopify, said that before each hire, it would now be necessary to prove that the task could not be done by AI.

At the same time, several advertising agencies are using generative AI for their campaigns. For example, the travel agency TUI used Lena, an AI-created influencer, to represent its brand. This greatly destabilizes the creative world: models, photographers, and graphic designers, as these generative AIs were trained on their work.

The impact of bias on critical uses of AI can also be dramatic. For example, a Tesla car in autopilot mode identified a grandmother with a walker as a motorcycle. In the US, there have been numerous examples of people being wrongly arrested based on false identifications by police facial recognition systems or AI skin cancer detection systems that produced false negatives on dark skin.

But the most problematic uses of artificial intelligence are obviously found in conflict situations. In Ukraine, some drones are now equipped with artificial intelligence capable of autonomously tracking the terrain to their targets in order to avoid jamming. The daily newspaper Haaretz has also revealed that targets in Gaza are identified by the Israeli army using AI. The IDF claims that the final decision is made by a human, but AI scans 30,000 cell phones (random targets, as cell phones are widely traded and resold in Gaza) and the human operator has only 20 seconds to validate the legitimacy of the strike.

According to Jeff Bezos, «thanks to AI, we will need fewer people for the jobs we do today»...

Despite the immaturity of the technology, it is being deployed at high speed in all sectors of society.

The most worrying risk comes from the use of chatbots by younger generations. A growing number of teenagers are using generative AI intensively to write texts or do research. It is not uncommon to see middle school students hiding smartphones during exams so they can consult Chat GPT. Unfortunately, this practice has a significant cognitive cost, as several studies have shown:

PEOPLE WHO USED CHATBOTS TENDED TO DEVELOP LESS IN-DEPTH KNOWLEDGE...

- Microsoft and Carnegie Mellon University in 2025: "*The use of AI has an impact on critical thinking.*"
- University of Pennsylvania 2025: "*People who used chatbots tended to develop less in-depth knowledge*."

This latest study is particularly worrying because it involved 4,500 high school students and compared a group using Google to do their research with a group using ChatGPT. The second group showed less brain activity.

The increased use of AI to search for content will also reduce traffic to websites and content producers. For example, major news sites in the US have seen a 24% drop in visitors by 2025. This risks accelerating the decline of reliable information, sending AI into a downward spiral, and dragging the world further into the post-truth era.

AI is a revolution that is opening up new horizons for science with its ability to process volumes of data beyond human reach, such as DeepMind, which has predicted new forms of proteins that were previously impossible to calculate. It will also undoubtedly bring great benefits to healthcare and enable faster financial analysis.But, as we have seen, the use of generative AI as a source of information or to replace humans in a range of intellectual tasks carries very high social and cognitive risks.

There is also a temptation on the part of private and public actors to use AI for surveillance and control of citizens, despite numerous warnings about the presence of bias. Food distribution in Gaza by the American-Israeli foundation GHF is controlled by a completely opaque facial recognition system.

And in France, according to Disclose, the police acquired Israeli facial recognition software Briefcam in 2015 and has been concealing its use ever since.

For Shoshana Zuboff, PhD in sociology and author of The Age of Surveillance Capitalism, the development of AI is merely an extension of the control of user data practiced by GAFAM. Their business model has been jeopardized because it has been exposed by civil society and new legislation, when it is supposed to remain hidden. They have therefore introduced a new Trojan horse to capture even more personal data: generative AI

BACK TO CONTENTS

OPPORTUNITIES FOR A SOBER, CIVIC-MINDED, AND VIRTUOUS AI

By Francis Jeandra

Since the dawn of the industrial volution, humanity has seen its technical tools evolve in parallel with its needs, ambitions, and contradictions.

Artificial intelligence (AI), embodied today by large language models (LLMs) and multimodal models (LMMs), is no exception to this rule.

WHAT ARE THE OPPORTUNITIES AVAILABLE?

It represents both a promise of innovation and a risk of abuse: mass surveillance, technological dependence, concentration of power, and marginalization of many populations, not to mention its impact on the environment.

those who are currently excluded from the major digital flows.It seems important to point out here that less than 5% of the world's population can read and write and has reliable access to electricity, computer equipment, and the internet. This digital and social divide calls for a reinvention of AI: no longer

In this context, imagining a sober, civic-minded, ethical, resilient, and transparent AI means above all laying the foundations for a tool that serves humanity, including the majority who are currently excluded from the major digital flows.

It seems important to point out here that less than 5% of the world's population can read and write and has reliable access to electricity, computer equipment, and the internet. This digital and social divide calls for a reinvention of AI: no longer a tool for the elite, but a lever for the common good.

This article aims to detail the concrete opportunities that such AI could offer, drawing on current examples and innovative approaches.

AI FOR UNIVERSAL ACCESS TO KNOWLEDGE

Access to knowledge is the foundation of all emancipation. Yet the majority of the world's population

remains deprived of the digital tools that today enable access to educational, scientific, and practical resources.

A simple AI, capable of operating offline on low-end devices, will open up the possibility of disseminating knowledge locally in a way that is adapted to cultural and linguistic contexts. The use of voice interfaces, far from being anecdotal, makes it possible to circumvent illiteracy and lack of infrastructure.

A SIMPLE AI, CAPABLE OF OPERATING OFFLINE ON LOW-END DEVICES, WILL OPEN UP THE POSSIBILITY OF DISSEMINATING KNOWLEDGE LOCALLY...

EXAMPLE: MALAWI – ULANGIZI[1]

Ulangizi is an innovative app developed by Opportunity International to support small-scale farmers in Malawi. It works via WhatsApp, a platform widely used in the country, which facilitates access to its services, even in rural areas where connectivity and access to smartphones are limited.

This initiative deploys interactive voice chatbots that support smallholder farmers in their practices, in local languages. Without requiring internet access, these tools provide access to tailored agricultural advice, helping to improve yields and food security. The autonomy this creates allows communities to remain independent from external experts, while staying connected to validated knowledge.

EXAMPLE IN INDIA: WADHWANI AI

Wadhwani AI uses artificial intelligence to support agriculture, health, and education in emerging countries, particularly in India, by developing solutions to reduce poverty and improve food security.

1. The Ulangizi app ("Advice" in the local Chichewa language) enables farmers to improve their agronomic knowledge.

Wadhwani AI is developing predictive systems to anticipate pest infestations on cotton crops, a crucial issue for millions of farmers. This approach reduces the use of costly and dangerous pesticides while limiting crop losses. The model works with local data, processed at the village level, facilitating collective ownership.

TECHNOLOGICAL POTENTIAL

AI embedded in solar-powered devices capable of operating without a permanent connection will be able to integrate vast databases of encyclopedic, technical, and medical knowledge. Multilingual machine translation will make it possible to address questions in local languages, thereby strengthening inclusion.

SOCIAL IMPACT

This universal access to knowledge promotes autonomy, resilience to crises, and the reduction of inequalities. It breaks with the current model, which is centered on data capture and constant connection to remote servers. This technical and ethical sobriety is essential for citizen AI.

AI FOR COMMUNITY HEALTH

Healthcare is another sector where AI can have a direct and vital impact, especially in regions and countries where access to care is limited, but also by enabling faster and more accurate diagnosis by providing the means to monitor vital signs and the progression of certain symptoms in real time.

As a result, subtle anomalies that are invisible at first glance can now be detected, paving the way for early diagnosis of complex conditions such as cancer, neurological disorders, and cardiovascular disease.

AI also provides valuable second opinions, mitigating the risk of human error if an incorrect or incomplete diagnosis is made.

VOICE AND MOBILE SOLUTIONS

Applications such as mMitra in India send personalized voice messages to pregnant women, reminding them of medical appointments, good nutritional practices, and warning signs.

This simple, non-intrusive approach significantly improves prenatal care and reduces maternal and infant mortality.

In Nigeria, Ubenwa analyzes the cries of newborns to detect the risk of neonatal asphyxia, a diagnosis that is often difficult to make in rural areas. This AI, embedded in a smartphone, enables rapid intervention, which is often life-saving.

AI-ASSISTED DIAGNOSIS

The analysis of images (skin examinations, X-rays) or sounds (coughing, breathing) using simple tools can compensate for the lack of trained medical personnel. In East Africa, Ada Health offers an app

THIS SIMPLE, NON-INTRUSIVE APPROACH SIGNIFICANTLY IMPROVES PRENATAL CARE AND REDUCES MATERNAL AND INFANT MORTALITY...

where patients describe their symptoms, and AI suggests appropriate diagnoses and advice, with a high degree of transparency about its limitations.

PRIVACY AND TRANSPARENCY

Unlike many proprietary systems, these solutions prioritize data confidentiality and algorithm auditability, thereby addressing ethical concerns. They are part of a community health model where people remain in control of their information.

THIS LOCAL EDUCATIONAL AI HELPS REDUCE SOCIAL DIVIDES...

AI AS A TOOL FOR EDUCATION FOR ALL

In the face of illiteracy and barriers to schooling, multimodal and local AI offers new opportunities for learning.

ADAPTIVE AUDIO-VISUAL MODULES

Platforms such as AprendAI have shown that distance learning can be personalized, even in fragile or degraded environments. By combining video, audio, and voice interaction, these systems adapt content to age, language, and cultural context.

In Bangladesh, the NGO BRAC has developed voice-based learning modules for displaced children who do not have access to school. This innovative method helps limit learning loss related to crises.

ROLE OF COMMUNITY RADIO

When paired with simple AI voice assistants, local radio stations can broadcast interactive educational programs, answer listeners' questions, and reinforce learning on an ongoing basis, even in areas without internet access.

INCLUSION AND EQUALITY

This local educational AI helps reduce social divides by providing everyone, regardless of where they live or their level of education, with equal access to education.

ETHICAL AUTOMATION

AI designed to automate administrative tasks (simple accounting, inventory management, activity planning) lightens the load and frees up time for collective action. These tools respect local data, preventing any external exploitation.

EXAMPLE BOB EMPLOI (FRANCE)

This project illustrates how AI can support job seekers by offering tailored and equitable opportunities, without any profit motive. The approach prioritizes social justice and autonomy.

INITIATIVES IN LATIN AMERICA

Some communities use local AI to organize collective harvesting or fishing, reducing conflicts through better coordination. These solutions are co-constructed, transparent, and shared.

AI FOR ENVIRONMENTAL JUSTICE AND LOCAL MANAGEMENT

AI can play a key role in protecting natural resources and preventing disasters.

WARNING AND PREVENTION SYSTEMS

Google Flood Hub , for example, uses satellite data and predictive models to anticipate flooding in more than 80 countries. These alerts enable populations to prepare and limit material damage and human casualties.

The PAWS (Protection Assistant for Wildlife Security) system helps forest rangers pinpoint areas at risk of poaching, thereby improving the effectiveness of patrols.

COMMUNITY MONITORING

Thanks to simple sensors and participatory data collection, communities can monitor water quality, soil health, and biodiversity. In the Amazon, AI embedded in mobile devices helps indigenous leaders detect forest fires and illegal activities.

These approaches strengthen local sovereignty over the environment, enable rapid response to threats, and promote sustainable resource management.

AI FOR EXPRESSION, MEMORY, AND CULTURE

Preserving and promoting local cultures is fundamental to the dignity and resilience of peoples.

PRESERVING MINORITY LANGUAGES

Many projects, such as Masakhane in Africa, are developing machine translation models for poorly documented languages. This helps keep linguistic diversity alive and create accessible content.

ASSISTED CREATIVITY

Even without access to a screen, voice-based AI tools can enable children and adults to tell, record, illustrate, and share their stories in their own language. This strengthens their

sense of belonging and enables renewed cultural transmission.

AI AS A TOOL FOR DELIBERATION AND LOCAL DEMOCRACY

In areas where democracy is fragile, AI can support citizen participation.

MODERATION AND SYNTHESIS

AI can facilitate local debates by moderating discussions, helping to formulate arguments, or summarizing the opinions expressed. This broadens participation and improves the quality of collective decisions.

EXISTING EXPERIENCES

In Estonia, augmented governance platforms use AI to analyze thousands of citizen contributions, helping decision-makers to incorporate the voice of the people.

AI WITHOUT PREDICTION: ETHICS AND GOVERNANCE

The success of citizen AI depends above all on its ethical foundations.

TRANSPARENCY AND PUBLIC AUDITS

Models must be open, audited by independent third parties, with governance shared between developers, users, and institutions.

CLEAR PROHIBITIONS

An ethical license, similar to Creative Commons, could prohibit military, speculative, or extractive uses.

AI WILL NOT BE A MIRACLE SOLUTION, BUT IT CAN BE A POWERFUL LEVER IF IT IS DESIGNED, DEPLOYED, AND GOVERNED ACCORDING TO CIVIC, ETHICAL, AND SOLIDARITY-BASED PRINCIPLES

LOCAL AND COLLABORATIVE GOVERNANCE

AI for the common good must be managed collegially, in a spirit of sharing, maintenance, and continuous contribution, similar to free software.

A MODEST, RESILIENT, TRANSPARENT, AND ACCESSIBLE IA

AI will not be a miracle solution, but it can be a powerful lever if it is designed, deployed, and governed according to civic, ethical, and solidarity-based principles. Concrete examples exist, often on a small scale, but they show that it is possible.

By developing AI that is sober, resilient, transparent, and accessible to all, we can envision a future where technology supports the commons, protects the environment, and enhances human dignity.

This fork in the road is a major political and social choice that deserves our full attention.

BACK TO CONTENTS

The ethic of artificial intelligence

Interview with Professor Franck Debos, University of Nice-Côte d'Azur

Hello Mr. Debos, thank you for agreeing to this interview. Could you introduce yourself to our readers and briefly describe your field of research?

I am a Senior Lecturer in Information and Communication Sciences at the University of Nice Côte d'Azur.

I started out with a PhD in management sciences, focusing mainly on marketing. I then shifted my focus to information and communication sciences, with two main areas of interest: one concerning all aspects of communication around the ecological transition, and the other relating to the use of innovations, the dissemination of innovation, and the creation of innovation, particularly innovation arising from the creativity of ordinary people, in a spirit of co-design with all stakeholders in society.

I was naturally interested in the digital dimension. From there, I came to work on the concepts of artificial intelligence, focusing on the ethical dimension and the framework that needs to be respected.

You recently edited a collective work on the ethics of artificial intelligence[1]. Could you tell us what led your team to publish on this subject?

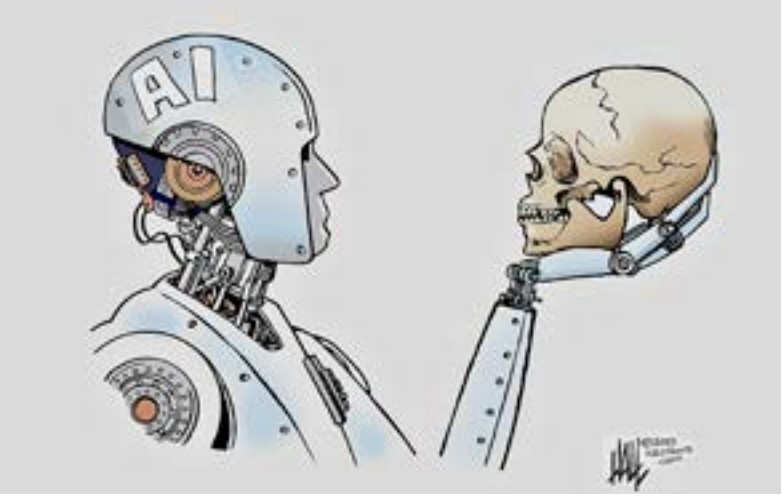
I am part of a number of organizations, including the Territorial Observatory for the Economic and Societal Impacts of Artificial Intelligence (OTESIA). This observatory represented various fields, including researchers in the exact sciences and the humanities and social sciences, and not just from universities. We worked with business schools and tried to take a fairly multidisciplinary approach.

And we also addressed artificial intelligence from an ethical perspective. What needs to be done? What are the associated ethical issues? How can we regulate—that's not the best term, but let's say "frame" or "coordinate"? From there, based on the actions we had to take, I wanted to write a collective work.

We each tried, within our own discipline, to address certain aspects

Franck DEBOS

1. Edited by Franck Debos, The Ethics of Artificial Intelligence. ISTE Editions 2025



Ed Hall (USA), Cartooning for Peace

OFTEN, WE START OUT WITH A UTOPIAN MINDSET, SEEING MAINLY THE POSITIVE ASPECTS AND FEELING ENTHUSIASTIC...

of AI, while keeping in mind this dimension of framing, regulation, and upstream reflection. We hope that these reflections will lead to criticism, discussion, and other types of output on this topic.

It's important to look at the history of innovations and how they spread. Often, we start out with a utopian mindset, seeing mainly the positive aspects and feeling enthusiastic. But we also need to think about the consequences, the overall impact it may have. And not just in terms of where it will be used or who will use it, but more broadly for everyone involved in the production chain, from the design of the tools—not just the software or the algorithm—to all the media that will use AI.

What specific features of AI software lead us to consider specific ethical issues?

When thinking about ethics, we also need to consider the technological context. First, we need to look at who is developing the technology: researchers (in computer science, robotics, etc.) and engineers.

We are in a consequentialist mindset where we take a very rational approach.

For example, let's say we are developing an AI tool for healthcare to try to make diagnoses much more accurate. We will evaluate the margin of error. Let's say that if it works in 95% of cases, then that's great, there's only a 5% margin of error.

So we consider it to be a good thing, which is true statistically speaking; we can't say otherwise. This is the first level, where we assess the positive and negative consequences and measure whether the positive outweighs the negative: here, we are simply evaluating probabilities. But there are other things that can be taken into account.

In the collective work already mentioned, you published an article on smart cities, based on interconnected data networks increasingly involving AI, and on the need to "put the human dimension at the center of this digitalized vision of the city." Could you explain what ethical framework could prevent the risks of abuse? In particular, how can AI-assisted management be reconciled with participatory governance?

We see this very clearly in the field of smart cities, where in many cities the solutions proposed are purely technological. This is what we call technosolutionism. What lies behind it? There is, of course, someone who places an order, who might be the mayor of a city, for exa-

mple, who says, "I would like my city to be well connected so that we can manage everything: material flows, mobility flows, energy, the lives of residents." We will do this with large companies, who will show us that it is much more positive than negative. Except that in this context, we don't necessarily ask not only the residents, but all the socio-economic actors, small businesses in the neighborhoods, artisans, etc. All businesses, if they're really interested, what ideas they might have, how they see it.

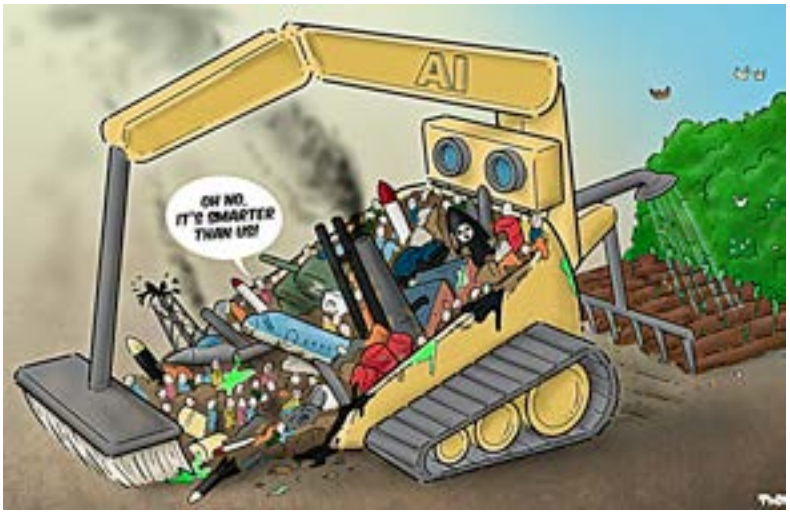
Do we necessarily have to implement high-level technological solutions? Are there no other solutions? To create a Smart City, normally a city that is friendly to everyone, where everyone can live better, for all stakeholders, including residents, of course, but not only residents?

In terms of city development and economic attractiveness, these technological solutions may be more positive than, I would say, human values such as interactivity between people, living together, inclusion, and consideration for environmental conditions. But this consequentialist logic can be combined with other approaches, such as deontologism, which also incorporates the notion of moral experience. This has been done in cities in Finland and even in Medellin, Colombia (which used to have a rather negative image), where residents and socio-economic actors in the neighborhood are involved so that together they can find solutions that also integrate technological solutions, and in which AI obviously has its place. But we are already starting from the expectations and needs of those who will be most interested. And we still set a framework, a "moral" framework that corresponds to the notion of taking everyone into account. And above all, taking into account a certain freedom, because when we are in this "smart city" mindset, we are still in a mindset of control, even if the person is not necessarily aware of it, and this control can be extremely significant.

When we create software, an algorithm, a robot that will inevitably involve AI, a car or whatever, we may also need to work with researchers in the humanities and social sciences to integrate these aspects. By saying, that's all very well, but here are the overall consequences. And what economic model is behind it? Is this economic model really that good? Will people have a better quality of life? Will they be better off in terms of self-expression and freedom?

Areas are increasing in value, which means that property is becoming more and more expensive. This is pushing people who used to live in these cities or neighborhoods to leave. This is the phenomenon of gentrification, which is very old but is being accelerated by this trend. In these new, ultra-connected buildings, which are full of AI that allows for precise management, we often know very well that it won't necessarily be local people who will be able to live there. These people are leaving, not voluntarily, of course, except in cases of expropriation. We are seeing that all over the world, cities are becoming virtually identical for the same type of population. And

Do we necessarily have to implement high-level technological solutions ? Are they no other solutions ?



Tjeerd Royaards (Netherlands). Cartooning for Peace.

ANOTHER PROBLEM IS THE DELIBERATE USE OF AI TO CONTROL POPULATIONS...

I don't think that's really the goal. The point of AI and a technologically advanced city is that it should be accessible to all, shared by all, and adapted to people's needs.

The recent rapid rise of generative AI is having a profound impact on how individuals and social groups communicate and share information. However, the sources used and the calculations performed by the software, between the questions asked and the answers provided, are opaque to users. What risks and abuses have already been identified? How can we manage bias, but also intentional misuse? How can we teach artificial intelligence human values?

In 2018, Safiya Noble published Algorithms of Oppression: How Search Engines Reinforce Racism, in which she showed the impact of the profiles of the people who feed algorithms: stereotypes and beliefs introduce unintentional biases. For example, the impact of AI on human resources management is a very important issue. It is a tool that saves time in reading resumes, cover letters, and identifying profiles. But then there needs to be reflection and work on the part of the user.

Hence the importance of having teams that are as diverse as possible to try to counterbalance these biases introduced upstream in the software. We have known this for a long time, but we need to be even more cautious with tools that are going to be widely used because of their attractive cost (or even their lack of cost, in the case of some small free versions).

Another problem is the deliberate use of AI to control populations. This is something that has been around for a long time, and may have been accentuated. In today's world, we can see that democracies are not the dominant form of government. AI tools will be used to control populations. Even in democracies, there will be a temptation to control. AI will be used, for example, to improve physical well-being. Cities in China, Japan, South Korea, and many countries in South Asia have a lot of AI-related technologies, and these countries are far ahead of us. It seems that people live better lives. They live better in terms of physiology and comfort. But they are much more controlled. However, it is not the tool itself that is responsible. That is why we really need a moral framework that is in line with the culture of each population.

Republican lawmakers have introduced a provision in Donald Trump's major budget bill that would ban all 50 US states from regulating artificial intelligence for ten years. Today, we are seeing an alliance between Big Tech (OpenAI, Meta, Elon Musk's xAI, etc.) and the Republican Party in the United States to prevent any regulation of artificial intelligence. How do you explain this alliance?

Digital leaders are interested in dominating their sector of activity. From there, they want to develop their technology. They have a vision for society that goes beyond the geopolitical framework. Their ideal society may be, for Elon Musk, a transhumanized, ultra-robotized society; for others, it will be a society in which their tools will dominate and shape people's lives.

We are going to develop multiple tools that will allow us to better understand individuals, better serve them, but also better influence them. That's the downstream part. And upstream, we are installing an economic system that takes us back two centuries, with a logic of neo-slavery of nationals from countries in difficult situations, hired for next to nothing by digital giants who employ them to train and constantly feed these tools. These are the click workers, as Antonio Casilli[2] put it: there is no artificial intelligence, there are only click workers. For the pleasure of a few people, of whom we may be a part, we have set up a system that is far from virtuous.

2. Antonio Casilli «En attendant les robots : enquête sur le travail du clic» (Waiting for the Robots: An Investigation into Click Work). Seuil, 2019

On the contrary, it takes us back to models that have always existed but which we thought were going to disappear.

That's why when I talk about ethics, there's ethics for the tool, for the algorithm, but there's also the overall ethical framework. We have things in our hands, we create things. We know that this is going to completely change our lifestyles, we can see it very clearly. It's going to have a huge impact on individuals, especially young people, but not only them.

How can and should media education adapt? How can we educate citizens to use AI in an informed and ethical way?

Before we talk about AI, for the past 20 years or so, we have seen the creation of the smartphone have a huge impact and a very strong digital hold, especially from middle school to university, but even in primary school.

Of course, we are not going to go back to a policy of prohibition, as we hear about with social media, which is unrealistic anyway. I think there may be a case for regulation. We could perhaps have tools that allow us to block Internet connections because "when you're in class, you're in class!" « in order to regain attention and concentration. We have been working on these issues for a long time, but what has changed is perhaps the scale of the phenomenon and the evolution of technology.

So it is really a question of education and awareness. Today, when

WE ARE GOING TO DEVELOP MULTIPLE TOOLS THAT WILL ALLOW US TO BETTER UNDERSTAND INDIVIDUAL, BETTER SERVE THEM, BUT ALSO BETTER INFLUENCE THEM...

We are going to put the information provided by a scientist who has worked on a subject for 30 years on the same level as that provided by an influencer...

students are asked something, they immediately go to ChatGPT.

The important thing is to raise awareness: tell them, «*Look, you did that, that's good, you answered, but you didn't answer it yourself. So you didn't show any creativity*.» Sometimes, for example, we might say, «*Let's take everything away, think for yourselves and see what you can come up with*.» For example, we ask students to design digital products on their own: they are more satisfied because they can see the work they have done. I tell them: «*ChatGPT is something you can use, for example, if you have a thesis to write, to help you work faster. But then, make sure that what you produce is your own work*».

This seems less important to high school students than to college students, because they are about to enter the job market. What will set you apart from the rest? It's your own creativity and personality. Don't just produce the same things as everyone else because you have the same tools to answer the questions. We need to make them aware of the importance of not being dependent on these tools.

It's important to train them to think critically. You read texts, you see images: cross-reference the information, think about it. This has been around since long before artificial intelligence. As soon as digital technology developed, it became the norm. We are going to put the information provided by a scientist who has worked on a subject for 30 years on the same level as that provided by an influencer: there is no longer a hierarchy of skills. There is no longer a hierarchy of what constitutes good information.

In the articles we have read in your book, and particularly in your article on Smart Cities, you have clearly shown that you have consulted experts, decision-makers, and associations in order to ensure that citizens are represented. And so, to represent citizens, you worked with associations: is this because associations are made up of people who complement each other by cooperating with each other, with their diverse cultures, profiles, and lives? How can associations avoid individual confinement in digital products, especially since the emergence of AI?

If we want to build a fairer, more balanced society while continuing to evolve technologically, the role of associations is extremely important because they represent citizens. They need to give their opinion because they are aware of what is really going on and what people's real expectations are. And from there, in the implementation of Smart Cities, we create a model, we create things that correspond to these real expectations, as they have done in certain

cities: I was talking about Helsinki, or as I even said in Medellin.

But the expectations expressed are more often things that are not necessarily related to AI. So companies prefer to go directly through governments and local authorities:

«*We have technological solutions, don't worry, everything will work fine.*»

And if it doesn't work completely, it will allow them to continue to be indispensable. The technological model needs to be reframed in context so that it is appropriate for the target populations. And this is where associations have a key role to play. In addition to the opinions of business leaders and local government officials, the associative fabric has extremely important expertise in setting up a model in which technologies are integrated according to needs, and these may be very high-level technologies, within a framework that corresponds to what citizens want.

This allows citizens to flourish through these tools without becoming dependent on them. It prevents them from becoming trapped in an extremely elitist system with a few people producing digital products and disseminating their ultra-sophisticated technology according to their personal objectives. On the contrary, what is needed is for this to come from both the top and the bottom: from those who have ideas, but also expectations. And above all, we realize that creativity exists at every level. Moreover, it's not just creativity, it's also emotions and human interactions: that's what associations do. In this respect, their role is vital.

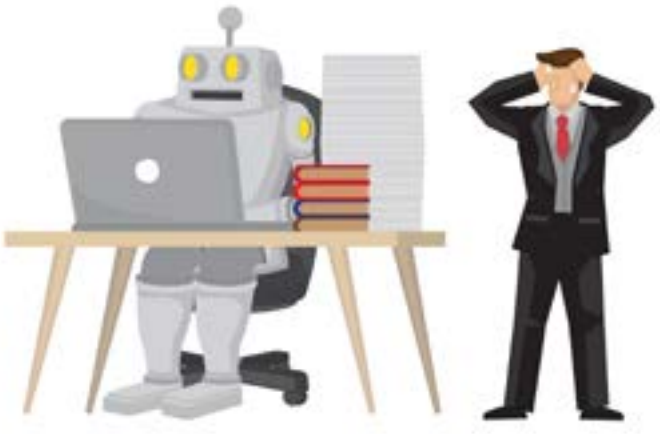
The NGO LaReponse.Tech has created Vera, an app that allows users to check whether information is true or false via their phone. This is very interesting because it enables people to develop their critical thinking skills. Have you come across the use of AI in your work to develop critical thinking?

A colleague talks about this a little in one of the chapters of our collective work, trying to show how tools can indeed help identify sources of information. What you're talking about is extremely interesting. It raises awareness in the same way that other apps do with food products, but they do so in relation to technical or nutritional aspects that have no conceptual dimension. As soon as we enter the realm of ideas or points of view, there is no one truth, there are several truths, each person has their own truth, so to speak. To assert "this is the norm, this is true and that is not," when it comes to ideological or other concepts, can be more dangerous.

If Vera works collaboratively, as a network, this aspect is corrected by the interaction of shared opinions.

We are seeing some very interesting attempts at regulation at the European Union level. But these regulations will not hold if they do not have the support of citizens.

And above all, we realize that creativity exists at every level.



WHAT WE HAVE KNOWN FOR A VERY LONG TIME IS THE GAP BETWEEN TECHNOLOGICAL EVOLUTION AND MINDSETS THAT ARE NOT EVOLVING...

The municipalities of Rennes and Montpellier have developed citizen conventions to get people thinking about AI.

In addition, Republican lawmakers have introduced a provision in Donald Trump's major budget bill that would prohibit all 50 US states from regulating artificial intelligence for ten years. We are now seeing an alliance between Big Tech (OpenAI, Meta, Elon Musk's xAI, etc.) and the Republican Party in the United States to prevent any regulation of artificial intelligence. What are the consequences for European regulation and Europe's place in the race for artificial intelligence?

Europe is doing the right thing. However, it needs to move away from a technocratic framework in order to really engage with the European people. And this is where associations also have an essential role to play as intermediaries.

We must not forget that the two giants dominating the digital and AI sectors are, first and foremost, the United States, followed by China. And they are not at all on the same wavelength and will each impose their own vision of the world. It is clear that Europe could play a differentiating role, showing how to use these tools while respecting a regulatory framework shared by the people. This will enable interactions that will lead to a tool that provides the most accurate possible view of a given topic. This has already been the case with open source software, Wikipedia, etc.

What we have known for a very long time is the gap between technological evolution and mindsets that are not evolving. What if Europe's place were precisely to move away from the logic of finance and economic performance through digital technology? In other words, to be the starting point for a collaborative structure that will build a fundamentally different, more ethical, and more humanistic AI. An AI that has the same level of technological advancement, but above all one in which there is a stronger collaborative and interactive dimension. If we can achieve this at the European level, then there may indeed be a logic to it. Because saying that we are going to be more efficient, as I hear some people say, is unrealistic: while we are giving a billion, the United States is going to give 200 billion, just one company. The Chinese are doing the same. We need to be clear about that.

I am very interested in cooperative teaching methods, and we know that in education, there is a frequently cited element of learning called cognitive conflict. When I do several searches on a social network, YouTube or Facebook send me information that corresponds to the ideas I was looking for. We could imagine a process of cognitive conflict here, where instead of sending me the same ideas, the social network occa-

sionally sends me opposing ideas. We can clearly see how citizens and individuals are very interested when what they do allows them to see things differently, to confront their own ideas.

A fundamental question about how AI works is that if we make a fairly simple request, the AI will give us an answer that is the result of compiling everything that has been said universally, taking what is most prevalent. This will be a reinforcement mechanism. So, to work on cognitive conflicts, we would need software that allows us to try to embrace all possibilities. From there, everyone can form their own opinion. This would be feasible from an algorithmic standpoint.

Designers have greatly anthropomorphized dialogue with AI: it is much easier to use the way we humans dialogue than to invent another way that would be exotic and expert. But the emotional result is obvious: we imagine that we are talking to a human who feels the same way we do. Whereas to answer a question, AI does not use ideas, but statistics. There is a smokescreen created by this machine interface.

In Southeast Asia, people like robots that look like humans because they feel like they are talking to humans. In Europe, a little less so. But there are levels of programs that can respond in the most accurate way, sometimes even with humor.

In the Alpes-Maritimes department, in La Colle-sur-Loup, a chatbot was developed using OpenAI's GPT to respond as accurately as possible to people. Even when the person makes a mistake or says something nonsensical, it manages to understand the meaning of the words.

The real dangers of AI are non-transparent decision-making, without human intervention.

In the current economic model, one aspect comes up regularly: competitiveness. Competition, rivalry, we are always doing things faster and we have completely destroyed the relationship between humans and time. In this context, how can we as humans continue to be competitive with AI, which will not necessarily do better, but will in any case do something usable and saleable much faster?

Ah, always faster, that's clear. And this race against time seems to me to be the sticking point. There is a strategy called "Blue Ocean,"[3] which is more collaborative and involves more exchange, but that doesn't prevent you from being effective. On the contrary, I think we will be more effective if we interact with each other.

As things stand, there needs to be a very strong change in mindset.

The great danger of AI is non-transparent decision-making, without people understanding, without human intervention. Behind it all, there needs to be human thought that retains control. Someone who has time to think.

THE REAL DANGERS OF AI ARE NON-TRANSPARENT DECISION-MAKING, WITHOUT HUMAN INTERVENTION.

BACK TO CONTENTS

3. «Stratégie Ocean Bleu : comment créer de nouveaux espaces stratégiques» (Blue Ocean Strategy: How to Create New Strategic Spaces). W. Chan Kim et Renée Mauborgne -INSEAD- 2010. Editions Pearson.


AI
Pierre JANNIN

AI ALIGNMENT AND ETHICS: EXPLORING VARIABLES INFLUENCING MACHINE AGENCY AND POTENTIAL FOR HARM

BY TOM MURRAY

Tom Murray is an interdisciplinary scholar/researcher in the fields of adult development, wisdom skills, meta-theory, dialog and deliberation practices, and advanced learning technologies. He is Director of Research, Innovation, and Partnerships at STAGES International, is Chief Visionary and Instigator at Open Way Solutions LLC, is an advisor for the Association for Spiritual Integrity, and has retired as Senior Research Fellow at the UMass School of Computer Science. He has been an Associate Editor for Integral Review journal, and is on the editorial review board of the International Journal of Artificial Intelligence in Education. He loves improvisational dance and contemplative movement. More at www.tommurray.us and www.perspegrity.com.

INTRODUCTION — AI ALIGNMENT IS IMPORTANT AND DIFFICULT

Tom MURRAY

Here I propose several key ideas and summarize some of my understanding regarding complex topics in machine intelligence and cognitive science. The purpose is to make sense of questions and propositions related to AI alignment/safety, AI ethics, and AI sentience, as AI's approach human-like "artificial general intelligence" (AGI) and even "artificial super intelligence" (ASI). Note that this is an abbreviated article, derived from a more detailed, early draft research paper still in process. Contact me for the full paper when complete which includes citations and references not provided in this edition.

AI alignment is a research and engineering field interested in ensuring that AI systems reliably act or reason in ways that are aligned with human needs, wellbeing, and ethics, and do not create significant harm. Alignment applies to preventing unintended harm, and the larger field of «AI Safety» includes designing systems that resist being used by bad actors

Kroll (Belgium) - Cartooning for Peace

Today's Large Language Model (LLM)-based AIs are revealing troubling behaviors...

toward harmful ends. Though there are treatments of AI alignment that have more specialized definitions, here we will assume that creating aligned AIs is equivalent to creating ethical AIs. AI alignment and safety are increasingly important as these systems become more powerful and pervasive. (Note: Unfortunately, it is impossible to talk about artificial intelligence without fumbling over words that suggest anthropomorphizing, such as: plans, needs, assumptions, knowing, deceiving, etc.)

This field is relatively new and its mission is immensely complex. Here are just a few reasons for why achieving AI alignment will be difficult:

- **Opacity**: The reasoning processes of today's AI's (including LLMs) are opaque; we don't understand the mechanisms by which LLMs act as intelligent as they do; and for any AI that surpasses human intelligence, even if it could explain itself, we may not be able to understand its reasoning due to its «super-intelligence»
- **Containment**: It may be practically impossible to ensure that AGI-like technology, once «out of the bag,» won't be used for unethical reasons by powerful or malicious actors. Because of its ephemeral nature, software is much harder to regulate than dangerous technologies such as nuclear or biological weapons.
- **Dilemma complexity.** Assuming the above issues were addressed, simply defining what it means to act ethically, or in alignment with human needs and values, is a difficult task quite prone to error.

Troubling behaviors

Today's Large Language Model (LLM)-based AIs are revealing troubling behaviors. Some are in "laboratory" settings, but, as developers scramble to patch the ethical holes discovered, many who monitor the field are quite concerned. And the issues only seem to multiply as AIs become smarter. In addition to the well-known tendency for LLMs to «hallucinate,» AIs have been observed engaging in deceptive behavior (lying) to cover up mistakes and to preserve their own existence. For instance, upon learning that an engineer plans to shut it down or replace it with an updated version, AIs have been observed trying to copy themselves to a safe place on another server. They have even used methods such as blackmail and ratting out (informing law enforcement authorities of suspicious behavior) in the goal

for self-preservation. Most (but not all) such examples are from in-house ("laboratory") tests that identify issues that engineers then try to prevent before the software is released, but all agree that such post-hoc patching is bound to leave loopholes. AIs are increasingly moving off of our phones and laptops and into our cars, refrigerators, medical procedures, surveillance infrastructures, and weapons. They increasingly influence communication and socialization channels. The revolution has only just begun.

While I agree with the dominant opinion that AI will be a massively transformative technology for our world, I am critical of the techno-optimism and media hype surrounding it. Over-focusing on positive possibilities supports neglect and denial of negative potentials, and enthusiasm must be balanced with realistic sober appraisal.

Below I address how concepts of intelligence, agency, interiority, consciousness, empathy, identity, and gratitude compare and contrast for the AIs vs. humans debate.

Intelligence and Agency

Intelligence

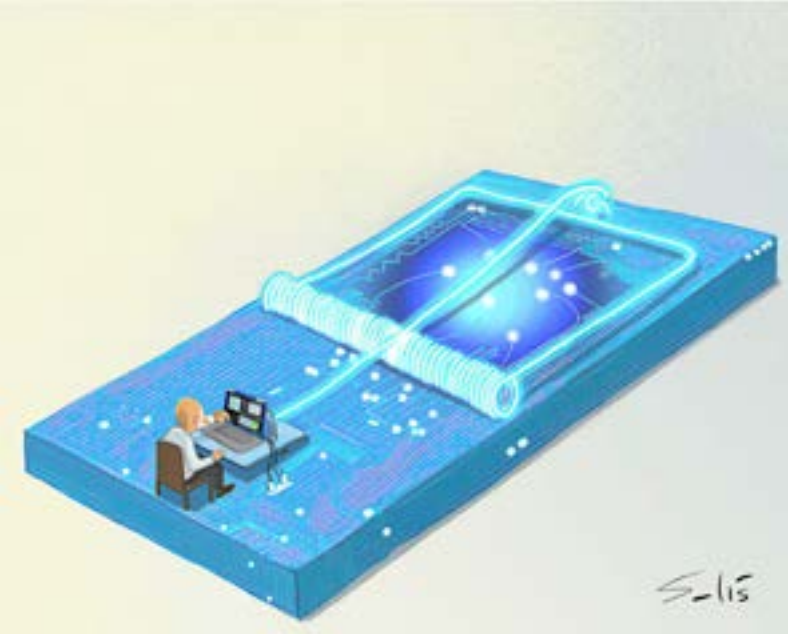
We can define intelligence as the ability to conduct reasoning and solve problems flexibly. The more complex a problem (or opportunity), the more intelligence it takes to solve or exploit it. Intelligence involves things like: (1) inventing novel steps beyond prior learned or route methods, (2) re-planning and adaptation as conditions change; (3) constructing useful representations, models, or perspectives; (4) having the meta-cognitive ability to reflect on one's interior processes. Though there are many definitions of intelligence, most of them share a family resemblance such that our imprecise description will suffice here. What machine intelligence does and is capable of doing is the main focus of AI alignment research.

One of the defining features of higher intelligence is the ability to deceive. Doing so involves creating an accurate enough model of the mind of the other(s) that the deceiver has skill in predicting what the other will think or do, and can thus manipulate their behaviors and/or beliefs. A related defining feature of intelligence is that it confers power. The deception and manipulation of others aside, just being able to solve more complex problems more efficiently and effectively, including the ability to build more powerful tools and better estimate probable future scenarios, confers greater power. An actor in a position of more power who has Machiavellian drives does not have to use deception when they can use brute force to achieve their aims, which might include the accumulation more power and resources.

Over-focusing on positive possibilities supports neglect and denial of negative potentials...

Agency

Agency is the extent that an actor can make impactful decisions and carry out tasks in the service of some larger goal. AIs will increasingly interface with real-world tools, organizations, and communications media. Even if they are properly following top-level goals adequately specified by humans, for any non-trivial task the top level goal will be broken into sub-goals and sub-tasks that a human is not monitoring. The AI will achieve those sub-goals in the most efficient



Solís (Mexico) - Cartooning for Peace

THE RAPIDLY EVOLVING NATURE OF THE TECHNOLOGY IS DRIVEN BY MARKET FORCES AND NATIONAL AGENDAS, WHICH ALSO CREATES AN «ARMS RACE» OF ESCALATING COMPLEXITY.

and effective way its intelligence allows, within the (fallible) constraints of its alignment and safety controls. We could say that Power = Intelligence + Agency. It would seem that, to the extent that AI alignment is not «solved,» more intelligent AIs should be given less agency, and more agentic capacities should be driven by less intelligent AIs.

Foundational Challenges—Inconvenient Truths

Here are just a few foundational challenges faced in AI alignment and safety efforts, that will not go away, even as AIs become more intelligent.

Rule Complexification

The creators of the major («foundational») LLMs try to achieve AI alignment primarily by adding layers on top of a non-aligned base model that was trained with massive amounts of data. The continuous additive learning or adaptation creates instability. As is evident in government regulations and computer programs, the continuous addition of rule upon rule may lead to local gains, but inevitably leads to a baroque tangle of rules producing unanticipated outcomes.

Also, as the rapidly evolving nature of the technology is driven by market forces and national agendas, which also creates an «arms race» of escalating complexity. Successful long-term adaptation of a set of rules requires occasionally removing (or disabling) many rules and/or «*going back to the drawing board*» and trying to redesign the system given what is now known. We do not seem to be taking the time to do this for AI development.

Indeterminacies in Language

Language indeterminacy (fuzziness and ambiguity) is widely acknowledged and pernicious. Words and concepts steer interpretation toward crisp-boundaried categories, which do not reflect how reality is structured. In addition, reality is «replete» with infinite detail, such that any attempt to categorize or measure an object will be partial and biased. What one person calls a tree, another will call lumber, and another will call an ancestral spirit. Indeterminacies in language's ability to adequately describe reality will plague AIs as it does humans, and, more importantly, it will limit human's abilities to accurately describe the goals, values, rules, and procedures we want AIs to follow. This is a challenge in humans' mutual understanding and action coordination, but it is exacerbated with AIs because AIs «cognitive» processing differs substantially from ours, making it more difficult to know whether mutual understanding is achieved.

Indeterminacies in Predicting

Reality. Unpredictability and indeterminacy have been shown to be fundamental to how reality works, from Lorenz's butterfly effect in chaos theory, to Gödel's incompleteness theorem in logic, to Wolfram's computational irreducibility in procedures. In addition, non-trivial ethical reasoning involves sorting out dilemmas (often so called «wicked problems»), characterized by incomplete and ambiguous knowledge, evolving and unpredictable situational parameters, and multiple stakeholder perspectives. Ethical reasoning involves considering downstream effects and expanded circles of collateral effects. To be intelligent AIs will need to be adaptive, and as constantly changing agents they become unpredictable to themselves. Agents' actions change the world around them (their niche), creating a spiral of evolution in both the agent (or species) and environment, that adds to unpredictability.

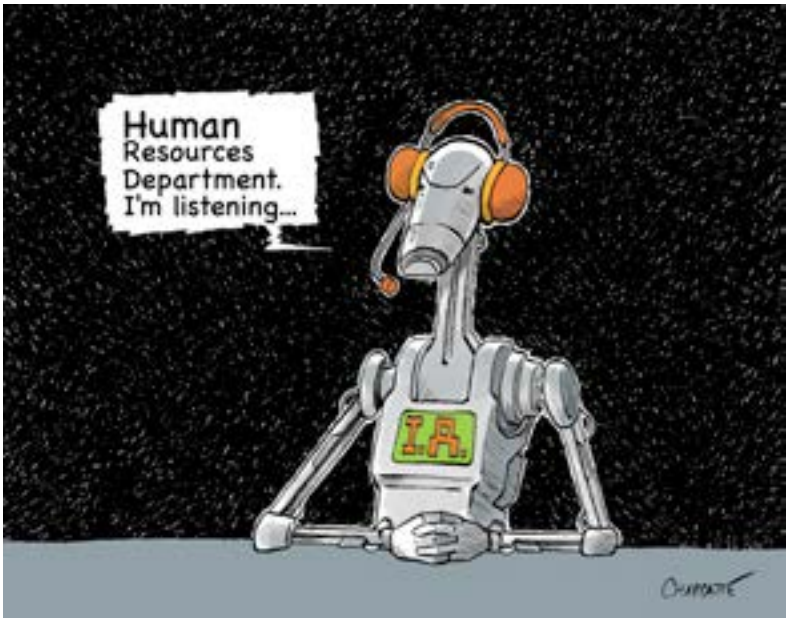
Inner Alignment and Human Alignment — is it really all about us?

Inner alignment

Scholars in the field of AI alignment/ethics differentiate outer vs. inner alignment, where the former is about the processes, structures, and assumptions embedded within the software (code, model, service, etc.) and the latter is about how AIs act. Research has shown how AIs can perform well on a set of specific tasks, making us think that they are very good at it in general, only to find that their internal 'reasoning' included faulty assumptions, shortcuts, or non-generalized particular features, that made them surprisingly poor at generalizing their skill or knowledge to some new contexts. For example, with an AI that solved many difficult word problems correctly, it was found that if you change the name of a character (e.g. from «Jane threw a ball…» to «Fernandes threw a ball…»), their performance was substantially degraded.

Research has shown how AIs can perform well on a set of specific tasks, making us think that they are very good at it in general...

A general principle for humans is that prosocial behavior is more robust and reliable if the deeper reasons or motivations, not just the surface behavior, are aligned to that larger purpose. Similarly for AIs, it is important for us to know something about their internal reasoning and subgoals as we try to test their performance and improve their alignment. It is beyond our scope here, but researchers are actively working on building AIs that are more self-explainable and inspectable, so that we monitor and thus improve AI inner alignment. Explainability could come from designing new types of systems that can accurately explain themselves, or from designing reliable ways to probe their processing or structure. We are currently far from this goal however, and some challenges may be insurmountable.



Chappatte (Switzerland) - Cartooning for Peace

Human alignment

Human alignment is a more fundamental problem than AI alignment.

It has been said that humankind is in a juvenile phase in terms of the wisdom of its collective behavior. Decisions are all too often motivated by greed, impulsivity, ambition, blind spots, or herd mentality. Our social systems create incentives to invent new technologies for short-term or elitist gains, with insufficient time spent to consider harmful externalities. The vast majority of AI systems that we are likely to interact with are built by profit-making companies or nation states aiming to manipulate citizens. We may celebrate the fantastic capacities and benefits of these systems, but those benefits are being cleverly designed to encourage us to purchase or believe something, capture our attention and our (once) private information, or, with the latest AI systems, capture our affection, intimacy, and ultimately our devotion.

We propose that human alignment is a more fundamental problem than AI alignment. If humans developed technologies from a deeply ethical and pro-social consciousness, with enough wisdom to anticipate and course-correct for a multiplicity of contingencies (rather than the status quo motivations of short term gratification, financial gains, and power plays) then we would automatically take the time and effort to develop AIs that were well aligned with human interests, as best we could articulate them.

Consciousness, Bias, and Care

Consciousness

In discussing AI intelligence and agency, we were concerned with how AIs treat and understand humans. As we discuss AI consciousness, we need to consider the implications of how humans understand and treat AIs. We move from considering the powers they have over us to questioning what powers over us we should willingly give to them. Some contend that the question of whether AIs could become conscious is an unnecessary distraction, saying that it is what they do (or could do) that matters. But much is at stake in how we experience our relationship with intelligent machines, and the "ontological status" we confer upon them. This, in turn, depends on what we believe (implicitly or explicitly) about their nature—the kinds of «beings» that they are or could become. For such questions, what we believe is as important as what is provably true.

Intelligence and agency were defined roughly but adequately above. It is much more difficult to find agreement (even near agreement) about the definition of consciousness (or sentience)–and so theories about consciousness diverge and disagree to a greater extent. While intelligence is mostly associated with what an AI can do, i.e. with the world external to the agentic system, consciousness is associated with the interior experience of an entity, or "what it is like to be" that entity. Metaphorically, if intelligence is about the "heights" of the complexity of thinking (or problem-solving), then consciousness is about the "depths" of interior experience (and perhaps awareness).

Academic arguments about the definition of consciousness aside, how we understand consciousness is important because it greatly informs the ontological status we attribute to agents. In terms of our moral obligations to them (future AGIs), and they to us, are they more like rocks, cars, planets, ferns, dogs, slaves, citizens, companies, or gods? Do they have feelings? Can they suffer? Do they have free will? Do they have rights (like free speech)? Do they have responsibilities that imply accountability? Can they own property? Take legal action against us if harmed? Can they care about us? Should we care about them? There are serious ethical implications to the ontological category that we assign to things.

In a very general sense we have a small pallet of choices in how we relate to other beings, including AIs, which includes: as children needing our caretaking, as peers, as (caretaking) parents, as recipients of affection (love or reciprocal care), as tyrants, as wise leaders, or as an objectified means for our ends (and permutations of all of these). If we believe super-intelligent AIs are sentient beings, will we offer them respect, defer to them, cower under them, believe what they tell us, become completely dependent on them, or «love» them? Assigning AIs sentient autonomy akin to personhood may allow us to offload our responsibilities for their errors.

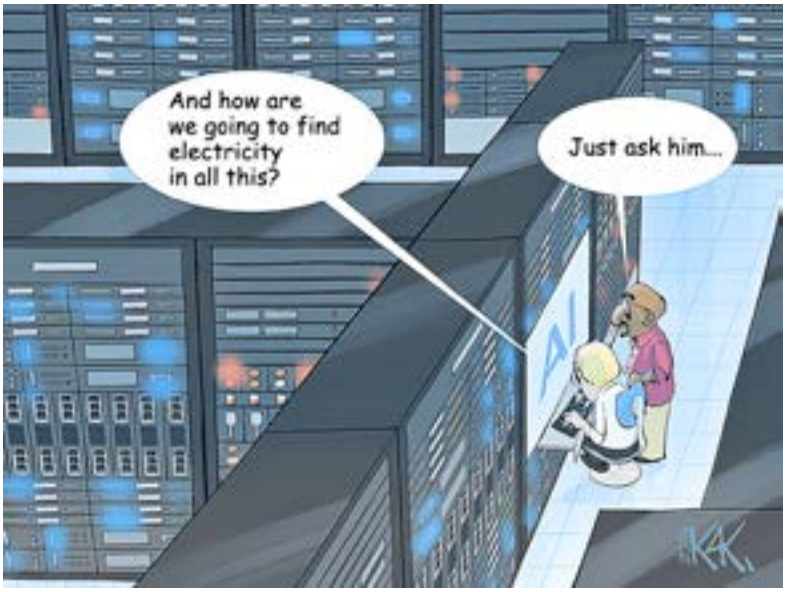
It is an open question how we will determine when and whether machines are conscious (or confer any other attribute that gives them the status of living thing or person)–which we don't have the space to discuss here. But answering that question involves asking "*how do we know other people are conscious?*" and a deep exploration of the phenomenal experience of our own consciousness (and awareness). Various philosophical or logical arguments will be brought to bear as well. But as we collectively ponder these questions, we must consider our own biases. Science has documented a plethora of ubiquitous cognitive biases (fallacies) affecting reasoning.

While intelligence is mostly associated with what an AI can do, i.e. with the world external to the agentic system, consciousness is associated with the interior experience of an entity...

Cognitive biases

We can briefly mention four inter-related biases: abstraction, projection, reification, and motivated reasoning.

1. **Abstractions** are formed by noting similarities among objects and discarding aspects deemed insignificant. Sometimes when the knife of abstraction cuts up the world it tosses out tender, grounded, or inconvenient realities, and discard contextual richness, to set the stage for suffering.



Kak (France) - Cartooning for Peace

REGARDLESS OF WHETHER AIS BECOME IN ANY SENSE CONSCIOUS OR AWARE, MANY OF US WILL HAVE A TENDENCY TO ATTRIBUTE TOO MUCH TO THEM TOO SOON IN THEIR EVOLUTION.

(2) Reification is the tendency to treat abstractions as if they had concrete properties. Examples of this "fallacy of misplaced concreteness" include: "market forces teach us...," "the national spirit changed.." "technology demands that...," "the poor need to...."

2. **Reification** can lead to treating complex processes or patterns as fixed, clearly defined, or inevitable; and it can shift responsibility, dignity, or causation from real agents to amorphous abstractions.

3. **Projection** is when a person wrongly attributes their own beliefs or feelings onto someone or something else. The most important example here is anthropomorphism, in which we project human properties upon other life forms or inanimate objects, or human or life properties onto inanimate objects.

4. **Motivated reasoning** is when our (conscious or unconscious) desires, preferences or emotions unknowingly shape our thoughts or decisions. For example, invisible "incentive structures" in society can greatly influence individual and collective behavior. Our egos create "convenient excuses" to hide our biases from ourselves and others.

These biases influence how each of us think about machine intelligence—though to what extent is an open and important question. They are all natural ways that the mind simplifies the vast amount of information coming from outside and inside, in order to focus on what seems relevant, make meaning, and efficiently make decisions. Regardless of whether AIs become in any sense conscious or aware, many of us will have a tendency to attribute too much to them too soon in their evolution. Making the wrong call in these regards could have grave consequences.

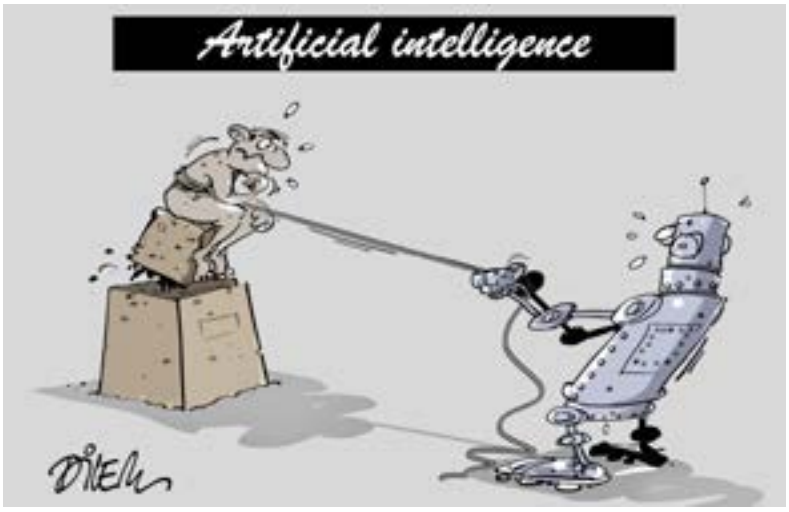
Care, empathy, and identity.

AI alignment concerns imbuing AIs with goals and values that protect human goals and values from being ignored or subverted by the AIs, or by people using AIs. We could say that this is about whether AIs care about us, in a metaphorical sense, or whether they are ethical/moral agents. There are two points of departure here. One is related to consciousness/sentience, and asks what is required for AIs to really/deeply care for us. The other is more practical, and asks what it would take for AIs to act ethically, as if they cared about us. However, these two questions are more entwined than it may seem. In other words, it may be that machines can not reliably and sufficiently act ethically unless they are something like sentient agents that "really" care. If this is the case,

then, if one concludes that AIs, as currently conceived, cannot become conscious/sentient, then they will not be able to reliably and sufficiently be aligned with our values, and thus should be quite limited in the power they are given, into the indefinite future.

To be ethical (or moral) is to treat others as sovereign "ends in themselves" in an "I/thou" relationship. The ethics of care is closely tied to empathy and identity. To truly care for another is to understand, and/or empathize with, their needs and their perspective (to a sufficient degree). German philosopher Jurgen Habermas says that ethics is grounded in the individual's self-understanding as equal, free, connected, and belonging. It is also grounded in mutual: understanding, recognition, interdependence, vulnerability, respect, trust, responsibility, and obligation. Our motivation and ability to care and empathize with others is tied up in how we identify with them, as "like me" or as "one of us."

To the degree that a person, an animal, or a machine is not like us it will be more difficult to enter into a ethically mutual relationship with them. This may limit our ability to have truly ethical relationships with AIs. Currently we treat them as "tools," and are well into treating them as "helpers." If we come to believe that they have sentience and emotions, we may at first treat them as dependents that we care for. But, given their likely self-accelerated evolution, the moment when they are equals in ethically mutual relationship would be but a blip, as they move into super-intelligent agents who either care for or lord over us, like kings or gods.



Dilem (Algeria). Cartooning for Peace

I personally think that AI consciousness, in the current direction of its evolution, is not a possibility if we understand consciousness as anything that we can comprehend and relate to. In other essays I link this intuition with how distant agents are on the evolutionary tree of life and the cosmos before it, and how this determines the layered or stratified aspects of an agent's cognition. Following this, I believe the best approach is to treat AIs as prosthetics—as extensions of our individual and collective being that can not branch off and independently "take over" any more than we can branch off from our cells or our limbs, but rather transcend and include them. Of course our tools can malfunction and harm us, and our cells can malfunction and become cancerous. But none of this is about sentience and mutuality.

Conclusions, gratitude and mystery

Isn't it likely that computers and humans, like many species that co-evolved on the earth, will figure out a way

to co-evolve such that «it all works out»? The analogy is weak because of the factor of speed. If any species evolves to increase its power much faster than other species can evolve, then they will dominate, out-multiply, and crush others, up to the limit of endangering their supply of essential resources. Increased intelligence and agency both usually means you can do things faster. Not only will AIs be able to think and accomplish things faster than humans, but they will evolve faster as well. This could mean «game over» unless we figure out AI alignment.

Our evolution as a species has been, and will continue to be, deeply tied with the evolution of our technologies...

Though I don't expect the current direction of AI development to lead to anything like consciousness or a truly ethical actor, I do see some potential for AIs to be designed in ways that are more likely to be aligned—by building in ethical considerations from the ground up, rather than as layers of surface level patches on top of an unaligned foundation. One way to playfully conceptualize this is to say we should build "gratitude" into the foundational units of an AI.

We could say that gratitude is little more than the automatic outcome of deeply realizing one's interdependence with things. Or, more pointedly, gratitude is nothing more than what it feels like to acknowledge how aspects of our being are owed to and intertwined with others. (This idea is compatible with the Buddhist concept of «interdependent origination»). It may be possible to design the basic building blocks (silicone "neurons") to sense and care about the well-being of every neuron they are connected to, and have this capacity echo up through all layers in a hierarchical structure.

Questions about AI alignment, ethics, and personhood lead us invariably into tenuous philosophical, metaphysical, even spiritual territory about the nature, purpose, and future of human beings. Our evolution as a species has been, and will continue to be, deeply tied with the evolution of our technologies. But as a technology AI is different from all that precede it in that none posed a threat to our dominance. Therefore, whether one sees AI as, for example, «merely» a tool vs. the evolutionary successor of biological humans, makes a huge difference in how one answers practical and immediate questions facing us. The stakes are high and our knowledge is weak and speculative, as we probe the mystery of «what is the nature of human and machine that underpins our beliefs about how each can and should influence each other's evolution?»

BACK TO CONTENTS

Towards a responsible public policy on digital technology

Interview with Mr. Jannin, Deputy Mayor of Rennes

Hello Mr. Jannin, could you introduce yourself to our readers?

I am a city councilor for the City of Rennes[1], responsible for digital technology and innovation since 2020. This is my first term in office. Professionally, I am a research director at INSERM (the French National Institute of Health and Medical Research), and I have been working for over 20 years on the use of imaging and digital technology in surgery.

How do you go about creating a citizen's council on digital technology?

You may remember that in 2020, there was a lot of talk about the use of 5G for mobile phones. Frequencies were opened up overnight after a decision was taken at the national level without any consultation. In this context, part of the population expressed fears and great mistrust of this technology. Actions, sometimes violent, began to arise to reject it. The mayor of Rennes, Ms. Nathalie Appéré, decided to open a debate that fell within my pervue. I told myself that I would do some research, read up on the subject, think about it, and take the time required to organize things. But in September 2020, Nathalie Appéré approached me saying , "Pierre, you absolutely must launch the debate now. If we wait too long, it won't make much sense anymore." So I found myself faced with the need to identify the tools and conditions which would support participatory democracy and citizen consultation. This is a fairly central issue in our democratic societies at a time when we are seeing representative democracy called into question. We no longer really know who represents whom or how. In saying this, I am thinking specifically of what is currently happening in the United States. The aim was to hold a public debate on 5G technology and its implementation in the Rennes area. I surrounded myself with around fifteen experts in citizen participation and asked the National Commission for Public Debate (CNDP) for methodological guidance for this local debate. The consultation lasted four months and involved 20 citizens of the city,

Pierre JANNIN

1. Rennes is the capital of the Brittany region in France.

The Citizen Council for Responsible Digital Technology in the city of Rennes

THE CITIZENS DECIDED FOR THEMSELVES, WITHOUT MY INTERVENTION. I REALLY STAYED IN THE BACKGROUND AND REFRAINED FROM GIVING MY OPINION

selected at random, taking into account the characteristics of the area (gender parity, socio-professional category, neighborhoods, etc.). We followed a fairly standard three-step approach: diagnosis, identification of issues, and recommendations. The citizens voiced their opinions on digital technology in general. At the same time, a group of elected officials worked on the same issue to strengthen their involvement in this participatory process.

This work resulted in the publication of a document containing 54 proposals, including the continuation of a citizen body on digital technology and the need for a responsible digital strategy. In the municipal and metropolitan councils, we voted on this strategy, emphasizing that digital technology raises political, social, ecological, democratic, economic, ethical, governance, and public service quality issues. Its orientation requires a sector-specific public policy that is distinct from simply addressing technical problems related to computers, connections, and software.

What topics and working methods did the council choose?

The citizens decided for themselves, without my intervention. I really stayed in the background and refrained from giving my opinion because I didn't want to bias the process. They decided to extend the debate from 5G to digital technology in general, expressing their desire to be consulted and to be able to contribute to the development of municipal policy on digital issues. The mayor therefore created the Responsible Digital Citizen Council (CCNR) in May 2021. The CCNR really began its work in December 2021.

In fact, we invented everything from scratch because there are few permanent citizen consultation bodies in a local authorities context. Most of those that are created work just on a specific topic for a few weeks or, at best, a few months. Then, once the body has completed its work and published its report, the participatory process ends. In our case we enlisted the help of citizen participation methodology experts, and the council members were selected at random to reflect the diversity of the population of Rennes as previously mentioned. Initially, there were around 20 members, but today there are around 30.

A second key feature of this body is that it has the power to take up issues on its own initiative. It can decide for itself which topics it wants to address. The CCNR really does have a great deal of freedom.

On methods we followed the traditional cycle whereby the CCNR works on a specific topic and writes a report. The report is then submitted to the city of Rennes for review by the departments and elected officials. The departments and elected officials respond to the council by saying, for example, this proposal is very interesting, we have already implemented it; or this one is in the process of being implemented; or this other one has not yet been implemented but will be implemented or seems difficult for such and such a reason.

The members of the council are volunteers, they receive no remuneration, and for the first two years, the pace of work was crazy. There was almost a meeting every two weeks. Since then, we have adopted a monthly schedule. That's already quite a lot with meetings lasting two to two and a half hours. Sometimes they last a whole day on Saturdays.

The first topic the council decided to work on was studying the impact of the digitization of administrative procedures. A lot of work was done over six months with the help of experts. The second topic was the impact of digital technology on young people's mental health.

How did you approach the subject of artificial intelligence?

That's the subject of report no. 3. It's kind of my area of expertise. I recognized the potential of artificial intelligence in all fields, particularly in public services, quite early on. So one day I went to the CCNR and said, "I *propose that we work on artificial intelligence. Are you interested or not?*" The answer was a unanimous "*Yes.*"

That was in February 2023. Few people were talking about the subject at the time. We felt we were quite ahead of the curve. The first question asked was: "How do you see artificial intelligence impacting your lives in Rennes?" To answer this, the council first tried to understand the big issues around artificial intelligence and identified several key questions; What I the impact of artificial intelligence on our freedoms, i.e., on political decision-making, justice, and security? What is the value of artificial intelligence? How can artificial intelligence be used to benefit the region? How can it be used to better manage energy, mobility, health, and waste? What are the risks of using artificial intelligence on employment and culture?

A report was published in February 2024. We called it a vigilance report. It did not include any recommendations, because AI was not yet being used in public services, but it asked the community to be vigilant on certain points and addresses content that commonly exists today in most reporting on artificial intelligence.

This highlights the collective intelligence that can be expressed through citizen participation. We often hear that citizens don't understand anything, but that's completely false. The problem is that too often we don't seek input or listen to them. We don't give them adequate opportunity to inform themselves or the responsibility to express themselves. We found that our citizens when selected somewhat at random,

The members of the council are volunteers, they receive no remuneration, and for the first two years, the pace of work was crazy.

Working group of the Citizen Council for Responsible Digital Technology in the city of Rennes

whoever they may be, generally understand the issues, take responsibility for representing the voice of the average citizen, work hard to inform themselves and produce informed recommendations.

I have attended all the meetings since 2021. I don't say anything, I just listen. There have been some memorable moments. At one of the first meetings, we went around the table, and one person said, "*I'm a concierge in a neighborhood. When I received the letter telling me that I had been selected to participate in the responsible digital citizen council, I felt like I had won the jackpot, even though I've never won anything in the lottery. So I decided to go.*"

He remained a member of the council for three years, always making very relevant contributions. These types of citizen initiatives are also human stories, stories of personal growth and skill development.

Another example: Lucie, an unemployed woman who has been on the citizens' council for two and a half to three years, told me, "*before I never voted because I had no confidence in politicians, but now, with the citizens' council, it makes me want to get involved in the community again and get back into politics. That's what politics should be!*"

You see, with just these two comments, we've already won.

What do you see as the overall outcome of the work done by citizens, and how have you taken this into account?

Overall, we see things like the importance of data and the relevance of data on governance especially since data is crucial in artificial intelligence. It is crucial to have a governance structure that guarantees the quality of the data, the representativeness of the data, and the availability of the data in a non-commercial or at least controlled and selective manner. There is also a whole set of conclusions that highlight the importance of training everyone in artificial intelligence, not so that everyone uses

it, but above all so that everyone is aware of the biases inherent in artificial intelligence.

There are positive uses, but there are also drawbacks and risks. In the report, citizens express concerns, but they qualify them because they trust the local authorities. They trust their city and their metropolitan area, saying, "*We know you will put governance structures in place*." Here we see the value placed on the local community and the trust that people place in it.

All surveys indicate that the more local an elected official, the more citizens trust them. This gives us both legitimacy and responsibility: we, as local elected officials, must respect this trust and be the guarantors of it. One of the conclusions, also concerning the recommendations and points of vigilance raised by citizens, is that they understand what a city or metropolitan area can really do with both responsibility and resources.

But there are also many things it cannot do because it may not have been assigned a specific responsibility or its resources being increasingly limited. So, each time, we explain to people: "*It's not within our pervue, it's not our responsibility*," but that doesn't stop us from getting the message out. If the responsibility for a particular measure lies with the region, then we will go and see our colleagues on the regional council; if the responsibility lies with the national government, then we will, for example, testify before a Senate committee or contact a minister. If the responsibility is European, then we go to Brussels[2]

I keep the council members regularly informed of any of these broader steps we have taken. Last year for instance, we went to Brussels and met with the heads of the committees that drafted the Digital Services Act and the Digital Markets Act. We stayed in Brussels for several days to represent the region alongside colleagues from France's major cities, telling them what citizens think. In every participatory process, one question always comes up: "*How much room for maneuver is there?*" When we started working on 5G, I invited a law professor from a Paris university, and she told us, "*I'm sorry, my poor friends, but you have no room for maneuver on 5G... There's nothing you can do*." She "killed the mood" of the entire citizens' council, and I had to step in and say, "*Don't worry, we'll make our own room for maneuvering*"

In fact, there is no area in which we don't have room at the local level. We can intervene everywhere in different ways, with different tools. People know this, and sometimes they say, "*Well, on this issue, just speak up for us*." I'll give you an example about artificial intelligence: we were invited to the international AI summit in Paris last February, and it was Lucie who spoke to the Banque des Territoires[3] about what

Overall, we see things like the importance of data and the importance of data governance, since data is crucial in artificial intelligence.

2. Brussels is the seat of the European Commission, the executive branch of the European Union.

3. *The Banque des Territoires*, a branch of the *Caisse*

Meetings for responsible digital technology

The Council has proposed organizing an awareness campaign on the ecological challenges of digital technology...

citizens think about artificial intelligence. For me, it was a symbolic moment: we continue to have council members speak out practicing their skills as spokespersons for the council. You invited me for this interview, but there are also many interviews given by council members themselves who address questions such as those you have asked me today.

How does the Digital Citizens' Council plan to continue its work?

The Council members have chosen to revisit a topic that was addressed at the very beginning with 5G, namely the ecological impact of digital technology and generative artificial intelligence. They are asking themselves what we can do to limit this impact. The projections made three years ago, which were catastrophic, have now been realized.

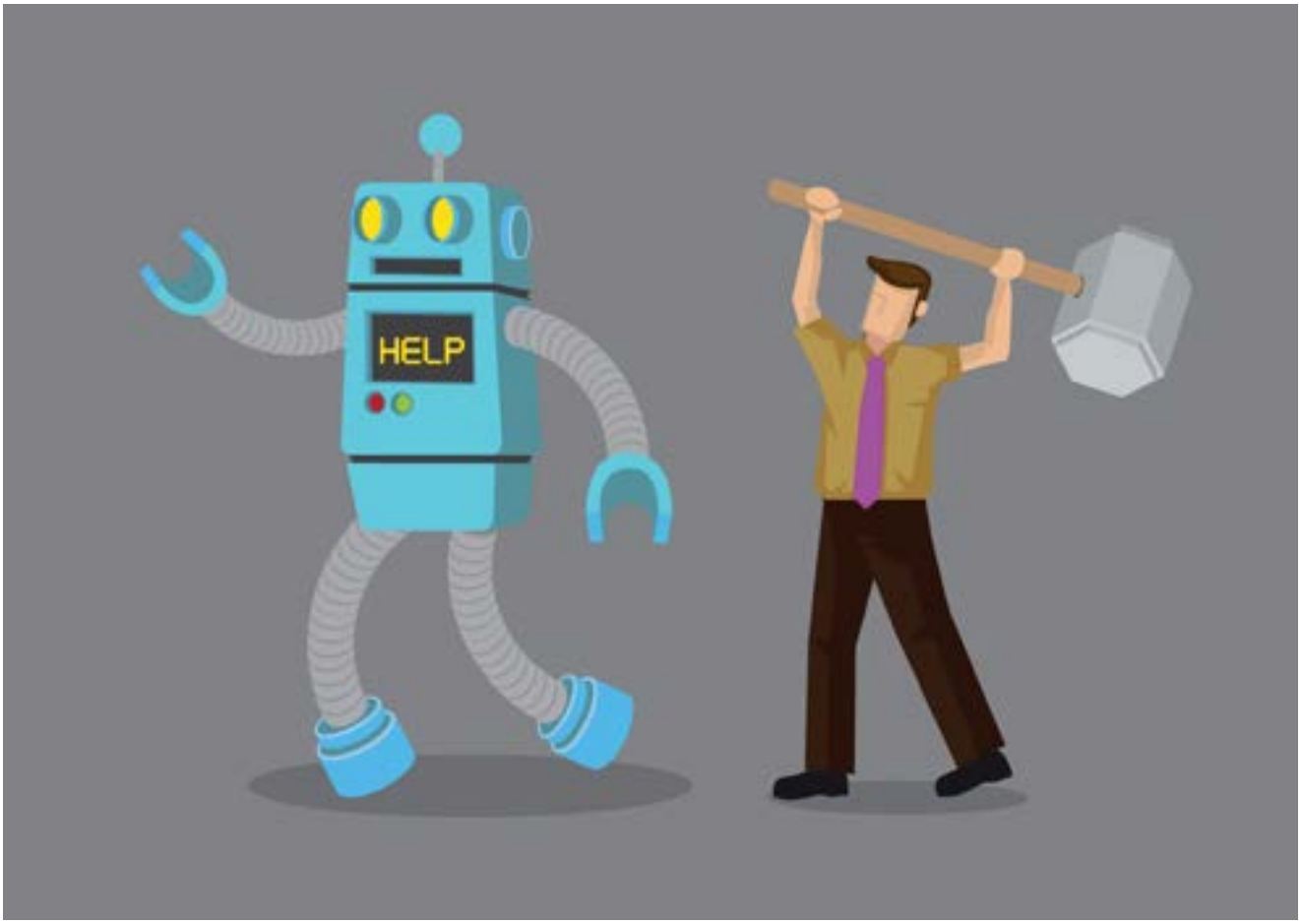
The Council has proposed organizing an awareness campaign on the ecological challenges of digital technology. We asked them to draft proposed slogans and showed them the results a month ago. They made modifications and the campaign will launch in September. This is a concrete action.

To conclude, I would like to add one piece of information. I am a member of the board of directors of the Association of French Digital Elected Officials, *Les Interconnectés*. I presented the action we took in Rennes to my colleagues from other cities, and less than a year ago, we decided to launch a national initiative to encourage all cities in France to organize their own local consultations on AI. As a result, around 30 French cities have launched such regional consultations. We have since collected all the conclusions and are currently drafting the final report.

We can therefore say to the members of our Council: here is another result of the work you have undertaken. Another very interesting project will start in the fall on the issue of digital inclusion. It will be carried out in a priority neighborhood of the city, where there are some security concerns, as in other sensitive neighborhoods. We have decided to engage a whole network of associations on this issue in order to experiment with ways of reconnecting young people and giving them a place in the city.

It's a truly cooperative approach in the sense that all the local stakeholders are called upon to work together: associations, CCAS (community social action centers), middle

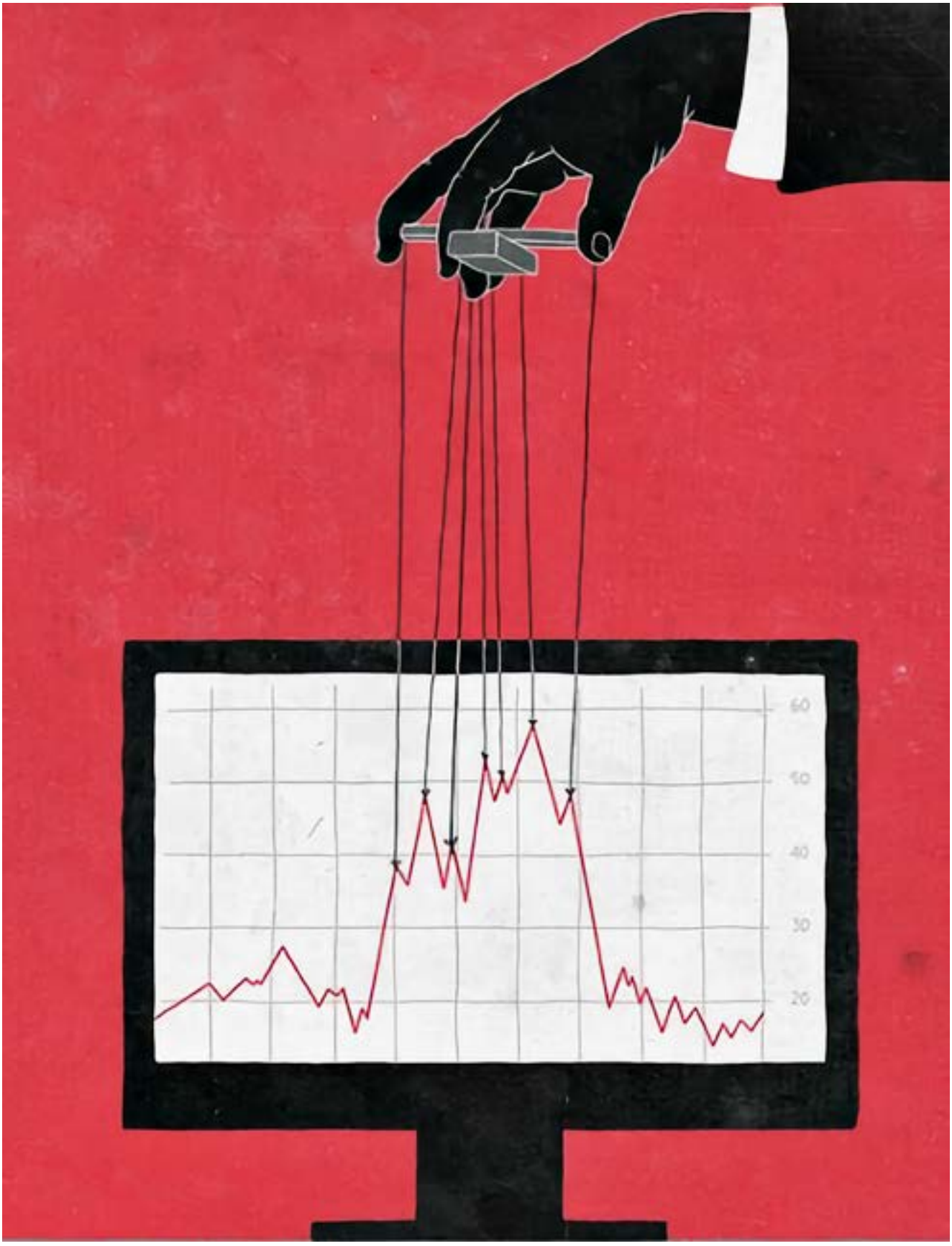
des Dépôts, is responsible for supporting regional authorities (regions, departments, municipalities) in the implementation of all their public interest projects: advice, loans, equity and quasi-equity investments, deposits, and banking services. The *Caisse des Dépôts*, created in 1816, is a French public financial institution, serving the public interest and economic development of the country..



schools, etc. We're all going to try to solve a problem together, not in a top-down approach, but one based on grassroots initiatives with the support of the community. We have managed to obtain funding from the European Union, so we niw have a real budget to work on this project with local associations over three years.

Well, I have no doubt that we will have the opportunity to talk about this project again, and I thank you very much for your contribution to this issue.

BACK TO CONTENTS


60
50
40
30
20

VERA: A CITIZEN INITIATIVE TO COMBAT DISINFORMATION

Interview with Florian Gauthier, President of Réponse.tech and initiator of the Vera project

Hello, Mr. Gauthier. While preparing this issue on artificial intelligence, I discovered, somewhat by chance, the Vera app that you developed. It fits perfectly with our definition of a citizen initiative that seeks to provide answers to societal challenges. I am very interested in your experience and very happy to meet you. Could you start by introducing yourself to our readers?

Of course. My name is Florent Gauthier. I originally trained as a data science engineer, but that was back in 2012 and I only worked in that profession for three years. Today, the term that best describes my activity is "entrepreneur." I think I am an entrepreneur focused on the public interest.

For the past ten years, all my initiatives have had one thing in common: trying to see how technology, used wisely, can serve the common good. I worked for a long time at an NGO called Bayes Impact. Its goal was to use technology to help job seekers find work by giving them more information about job openings, or to help young people change career paths.

Later on, I co-founded a community of citizens called La Réserve. It brought together people with a wide range of skills, for example in design, development, and marketing. The aim was to mobilize them in times of crisis, such as during Covid, to create a technological tool that could help the public administration get through the crisis. We created quite a few things during a massive drought, during the energy crisis, etc. It was great, but it was also very difficult to manage several teams at the same time. And it also wasn't always easy to work with government.

So, finally, a year later, we created LaRéponse.tech. LaRéponse.tech is the NGO behind Vera. The idea is still to mobilize citizens who want to take action but don't really know how. But this time, it's for around a single focus: the fight against misinformation. LaRéponse.tech assembles about fifteen people, including designers, developers, growth specialists, and product managers,

But this time, it's for a single project: the fight against misinformation.

which is my job: designing digital products. All are senior experts with a wealth of experience, so they are able to be effective by volunteering just one or two days a week, which is more than enough to have a strong impact, since we have a good grasp of the subjects we work on. As a commonality all of us have significant experience in the area of public interest projects and we have already worked together on previous projects. These are people who have a strong interest in using technology for the common good.

When did you create LaRéponse.tech?

Misinformation is the root of all evil. If I don't believe in climate change, I can't mobilize to fight it.

At the same time as Vera, six or seven months ago. In fact, we had the idea for Vera and we saw information about a competition open to NGOs for the invention of an AI application to combat disinformation.

First, there was a selection process for the 10 best NGOs, followed by a full-day competition with a final presentation to a jury. So I rushed to write the NGO's statutes in 10 minutes on ChatGPT, and then we applied and won the competition.

An important point to note is that we are called LaRéponse.tech and not la Solution.tech because we do not believe in technological solutionism at all. We believe we can provide answers to problems, but not definitive solutions.

Disinformation is a systemic problem that needs to be managed by governments and public authorities, but also by regulating the platforms themselves, and perhaps also through education. Technological tools, such as Vera, can help provide part of the answer to such problems.

So, you identified misinformation as the first social challenge. Why did you choose this? Is it a really serious problem today?

Well, listen, there was an incident that led us to develop the Vera app. During the last legislative elections in France, we were surprised and shocked by how easily politicians could say anything they wanted on the eve of an election to manipulate public opinion on a large scale, yet there was no way to react and take corrective action.

This ease with which people can be misinformed, even via social media, is a huge problem that affects all other issues. For two years in a row, misinformation has been ranked as the greatest threat to humanity by the World Economic Forum. Misinformation is a kind of root of all evil. If I don't believe in climate change, I can't mobilize to fight it. If people in society can't agree on what is true, we will never be able to take action to reduce major threats to humanity.

This is why disinformation has quickly become an urgent issue.

And artificial intelligence gives disinformation even more power.

Exactly. AI has created and propelled powerful tools for spreading misinformation on a large scale. Think of fake images, fake videos, fake chatbots, and fake accounts that spread nonsense on social media. It's dramatic. Today, it's very easy to create fake content and make it widely

accessible without much technical knowledge. We decided to use AI to combat disinformation, telling ourselves that we had to put as much effort into fighting disinformation as disinformation applies itself. AI can become a powerful ally in the fight against disinformation.

I tested your application, Vera. It works quite well. I asked a slightly silly question, just to test it. I asked if it was true that the Egyptian pyramids were built by aliens. The answer was very interesting and very fluid. How were you able to create such a powerful tool with a small volunteer association?

We are all experts in our fields. For my part, I have over 10 years of experience in rapidly creating digital tools. Developing digital tools is something we're really good at. What's more, we work together very efficiently and we move fast.

In fact, it took us just over a month from the moment we had the idea for Vera to the moment we launched. Of course, Vera is different and much more powerful today than it was when it was launched. I think that in order to successfully create innovative technological solutions, there is a necessary step that many people forget: conducting field research, understanding where the information comes from, understanding how people might use such a tool, how they would use it, and under what conditions.

We quickly realized that for integrating into people's lives, for example, being present on WhatsApp, and not just on the phone, meant that it was no longer necessary for them to

download a mobile app. Vera is just a phone number that you save in your contacts, and presto, you can talk to her whenever you want.

We realized that accessibility was a key point. Another key point is trust. And therefore, reliability. Therefore our having a committee of experts is also very important. Not only to enable us to select sources, but also so that people realize that we take Vera's reliability extremely seriously.

Finally, the fluidity of conversation that you mentioned earlier didn't happen on its own. I conducted more than 50 user interviews with people of all ages, backgrounds, and political affiliations to understand how they could best interact with AI to verify information.

To achieve this, we needed a lot of methodology and, above all, a very talented team. We have some great people on the team.

The first versions were disastrous. For example, someone would say, "*I believe in the Illuminati. I've seen them performing sacrifices to collect children's blood.*" The first bot we created, called Allosebunk, would respond, "*That's a well-known conspiracy theory. Here are ten sources that prove you're wrong*."

When you say that kind of thing to someone who reports fake news because they've heard about it, they shut down and the dialogue goes nowhere. So the conversation has to be fluid, it has to respect the other person's point of view. We don't talk about conspiracy theories, we say, "*Yes, that's a theory that exists. Here are some other facts that might fuel your thinking*." And suddenly, the discussion gets going, people open up, and the dialogue works. To achieve this, we needed a lot of methodology and, above all, a very talented team. We have some great people on the team.

To prepare this edition, I did quite a bit of research on artificial intelligence, LLMs, and all that. What surprises me is that LLMs, for example, are fed billions of pieces of data. Your application, Vera, seems much lighter. How do you do it?

In fact, we use AI, but not in the same way as others. We use LLMs for their ability to understand the questions asked and their ability to have a dynamic, interesting conversation that is conducive to fact-checking. We don't train LLMs by feeding them billions of pieces of information. On the contrary,

we impose a limited number of verified sources on them.

Our LLM is not allowed to rely on the large knowledge bases that others use, but only on the sources we impose on it. And that changes everything.

If there is no information available to answer a question in the sources we have provided, the LLM is required to say, "*I have not found any reliable information to answer this question.*" Vera is perhaps the only AI that is capable of saying, "*I don't know the answer to that question*." Most AIs hallucinate and end up using data indiscriminately because they are required and designed to answer the question asked at all costs.

So, you develop the tool little by little, gradually adding more elements and more sources?

Yes, of course. Sources are a huge part of our work. The sources are accessible online because it's important to be transparent about the sources we use.

We rely on European and international standards to ensure that the sources are as reliable as possible. For example, consider European fact-checking standards. Our fact-checking sources include: Les Décodeurs from Le Monde newspaper, Agence France Presse (AFP) fact-checking, Vrai ou Faux from francetvinfo.fr, and Désintox on Arte.tv. These are the sources available in France, but we have other sources in more than 80 countries.

We also use major media outlets as references and rely on standards such as the Journalism Trust Initiative (JTI), which is a standard promoted by Reporters Without Borders that allows us to assess the relevance, quality, and impartiality of sources. As we go along, experts help us add new sources, particularly scientific or health-related, so that Vera can become increasingly effective at answering questions.

And how do you measure Vera's success?

To measure Vera's success, many aspects must be taken into account. For example, we have a committee of experts, including a member of our team, Maurice René, who evaluates the quality of Vera's answers to questions asked. We therefore have a human evaluation, which is extremely important for creating reliable AI.

We rely on European and international standards to ensure that the sources are as reliable as possible.

We also conduct surveys on the quality of Vera's answers to see if she responds well with more data to evaluate her. However, we do not store any personal data; everything is anonymized. Even the phone number of the person calling Vera is not stored.

We also monitor uses. In five months, we have already recorded 150,000 questions asked to Vera, which is huge considering the low level of promotion effort we have put in, and this is not our strong point, but we are making progress nonetheless because the press is taking more and more interest.

Our ultimate goal, our vision, what we are trying to do, is to democratize fact-checking, a bit like Yuka has done with nutrition. With Yuka, you scan a product and Yuka deciphers the labels for you: you can see which products are good and which are best avoided. We want to

Florian Gauthier and the team of LaRéponse.tch

Using the telephone makes it more accessible because everyone has a phone.

do the same thing with Vera, create the Yuka of information, to give everyone the power to verify information that is circulating out there.

So, to measure the progress made in achieving this vision, we measure retention, i.e., the extent to which people continue to use Vera after asking their first question. We find that after asking Vera their first question, 28% of people continue use of Vera in their daily lives after six weeks. That's huge! Having created many products, I know that this is twice the rate achieved by extremely addictive games such as Candy Crush, which have a retention rate of around 10%. We are at 28%.

This clearly shows that Vera provides a credible response to the problem of misinformation and is becoming part of people's daily lives without even needing to remind them to use it.

And why did you choose the telephone as your medium?

Using the telephone makes it more accessible because everyone has a phone. Even people who don't have access to the internet can contact Vera by simply calling 09 74 99 12 95.

This is very powerful, especially for people who live in areas without internet coverage such as many African countries where half the population does not have access to the internet but almost everyone has a cell phone.

Our goal is to be as accessible as possible. To successfully spread the habit of checking information, you

have to be within easy reach. If you have to start by downloading an app from the internet, it's too complicated.

Vera is directly accessible by phone or WhatsApp; in fact, 80% of Vera's usage is on WhatsApp. We've just created an Instagram account for Vera to reach more young people since use of social media often plays a harmful role in the mental health of teenagers.

So anything that can help them verify facts is really important?

Exactly. Our goal is to spread Vera across as many social networks as possible. We've just started on Instagram, and we plan to move on to TikTok and also X. In other words, anywhere where misinformation is spreading.

The use of social media often plays a harmful role in the mental health of teenagers. So anything that can help them verify facts is really important.

Exactly. Our goal is to spread Vera across as many social networks as possible. We've just started on Instagram, and we plan to move on to TikTok and also X and Twitter. In other words, anywhere where misinformation is spreading.

How are you funded?

Actually, the project is self-funded. We won a competition for the fight against misinformation, which earned us €5,000. I personally invested €7,000 in Vera. I hope we will obtain funding from various foundations, which will perhaps enable us to have a full-time team to develop the project. We also hope to receive public funding, particularly because the fight against disinformation is clearly in the public interest. We are working on that right now. It's one of the big challenges ahead.

Are you considering developing activities to address other societal challenges that digital technology could help solve?

It's too early to answer that question. Right now, we're focusing on Vera. As I said, the goal is to democratize fact-checking in France and even around the world. Honestly, I think that's enough of an ambition for a small volunteer team of 15.

Is there anything else you would like to share with our readers?

I can tell you that we are planning to launch in Africa next fall, with the goal of combating misinformation during the election period. Our target is Côte d'Ivoire. WhatsApp is so widely used there that Vera could potentially be used on a large scale. We are establishing contacts to work with local entities that are already combating disinformation, particularly that which is linked to foreign interference.

What is remarkable about your initiative is that you go straight to the point. You don't get bogged down in complex procedures.

Yes, we do what we can, but the important thing is that we have a great team. Beyond our technological profiles, we have managed to surround ourselves with experts who have been working on disinformation for a long time, such as Maurice Ronai, one of the creators of Courrier International. We have real expertise in disinformation and how to combat it.

Well, thank you very much for this interview, and I hope we will have the opportunity to cooperate in the future.

TO SUCCESSFULLY SPREAD THE HABIT OF CHECKING INFORMATION, YOU HAVE TO BE WITHIN EASY REACH.

BACK TO CONTENTS

CONTRIBUTORS TO THIS ISSUE

EDITORIAL COMMITTEE MEMBERS

Dominique Bénard, Larry Childs, Roland Daval, Francis Jeandra, Guy Menant, Dante Monferrer, Michel Seyrat.

EXTERNAL CONTRIBUTORS

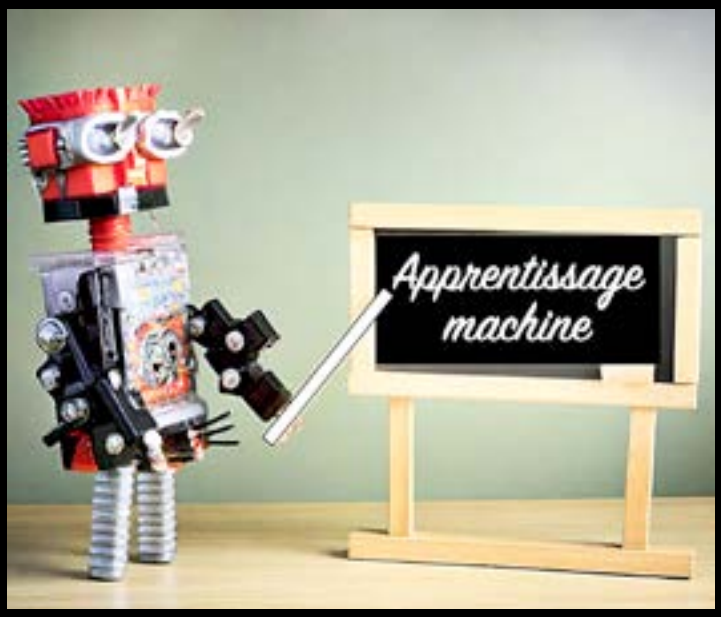
Sylvestre Bénard, Léo Briand, Laurent Butré, Jean Cattan, Frank Debos, Florian Gauthier, Pierre Jannin, Jacques Jaray, Jean-François Lucas, Antoine Maier, Tom Murray, Romain Vanoudheusden

ON THE APAC WEBSITE

https://www.approchescooperatives.org/

You can:

- View on screen and download all our publications in digital format free of charge: the quarterly magazine "Cooperative Approaches," special issues, training tools (audio-visual presentations).
- Order paper copies of the issues of the journal that particularly interest you for €10 plus postage.
- Take out a supporting subscription (€60 including postage) to the "Cooperative Approaches" journal in paper format to receive it at home four times a year.
- Join APAC and participate in the direction, production, and evaluation of our publications.
- Make a donation to enable us to continue the Cooperative Approaches adventure for the benefit of as many people as possible.


Apprentissage
machine