Member-only story

The Neuron That Wanted to Be God

7,000 words to tell an 80-year-old story

Alberto Romero Follow 27 min read · 1 day ago

I.

When scientists Warren McCullough and Walter Pitts invented the first mathematical model of the neuron in 1943, they couldn't have imagined that eighty years later, humanity would be pouring hundreds of billions of dollars into their idea. But here we are, and they are not, so they cannot tell us just how patently absurd the current artificial intelligence frenzy has become.

If you take a look at the modern pipeline of how AI models are created — from the first line of code some unknown developer writes to the moment Sam Altman takes the stage and claims "With this new model, I did feel the AGI" — there are various points at which one can seriously argue that "ok, this should have been done some other way." (Maybe it's not something to be proud of that the next generation of AI models needs a quadrillion tokens of data — that is, 1,000,000,000,000,000 as many — to increment their performance half a percentage point in some unknown benchmark test.)

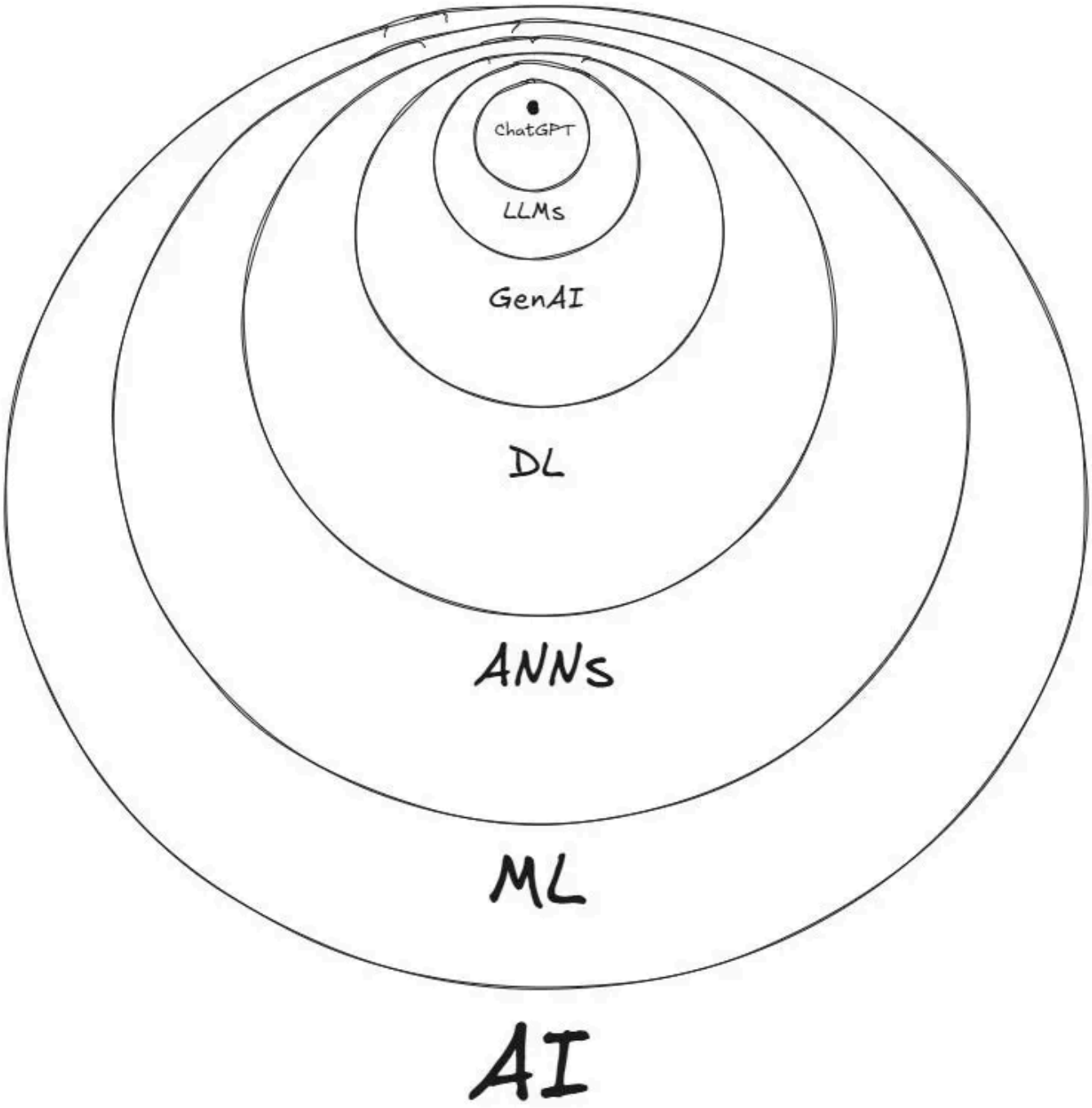
However, it's genuinely surprising that you'd have to go *all the way back to the inception of the field* to find the first controversial decision that never received sufficient care or consideration — not even after more powerful computers and more money and more data tokens were devoted to the cause.

You could wonder why AI labs are trying to scale large language models (LLMs) to the size of the human brain (in the ballpark of a quadrillion synapses), or why they think that combining deep learning (DL) and reinforcement learning (RL) is sufficient (that's the current pre-training + post-training strategy) or even whether artificial neural networks (ANNs) is truly the only paradigm they need, and still you wouldn't be pointing to the original sin of the field: accepting the simplicity of McCullough and Pitts neuron model as a faithful representation of its biological counterpart, which has proven to be orders of magnitude more complex functionally and structurally.

AI people, for all the love they profess for brain-inspired metaphors, never revisited that one. It might be too late to backtrack now, but it's never too late to call out the naivety and folly of those who have humanity's fate in their hands. This might sound dramatic, but that's because you haven't read this other article of mine: "They Are Sacrificing the Economy on the Altar of a Machine God."

We need to understand how we got here from McCullough and Pitts' work, and the first step is to clarify some common misconceptions. Many of you have come to think of ChatGPT or generative AI as a default label to encapsulate the entirety of AI. I don't blame you; the industry has made a great effort to push this synecdoche (the part is the whole): if the most popular sub-sub-sub-field can gather so much investment and interest, why

not let this mischaracterization leak into history books and the public opinion uncorrected? It's a marketing bargain. But, if we were to conceive a taxonomy of the AI field — such a daunting task that not even Wikipedia has managed to do well — we should reject this fallacy and instead draw a Venn diagram like this (experts will forgive any inaccuracies):



A toy taxonomy of AI down to ChatGPT. If I were epistemically rigorous, I'd have to define classification criteria first and then account for all the exceptions. But neither have I time nor you the attention span for that.

Ok, so that was easier than I thought: the AI field is basically a bunch of nested circles, not even a Venn diagram in the strict sense because none of them intersect with one another. So we have that artificial intelligence contains machine learning, contains artificial neural networks, contains deep learning, contains generative AI, contains large language models, contains ChatGPT. Or, in mathematical notation (quote-ready): AI ⊃ ML ⊃ ANNs ⊃ DL ⊃ GenAI ⊃ LLMs ⊃ ChatGPT. (Show this to your friends, and they will think you're either 1) an intellectual genius or 2) an insufferable nerd.)

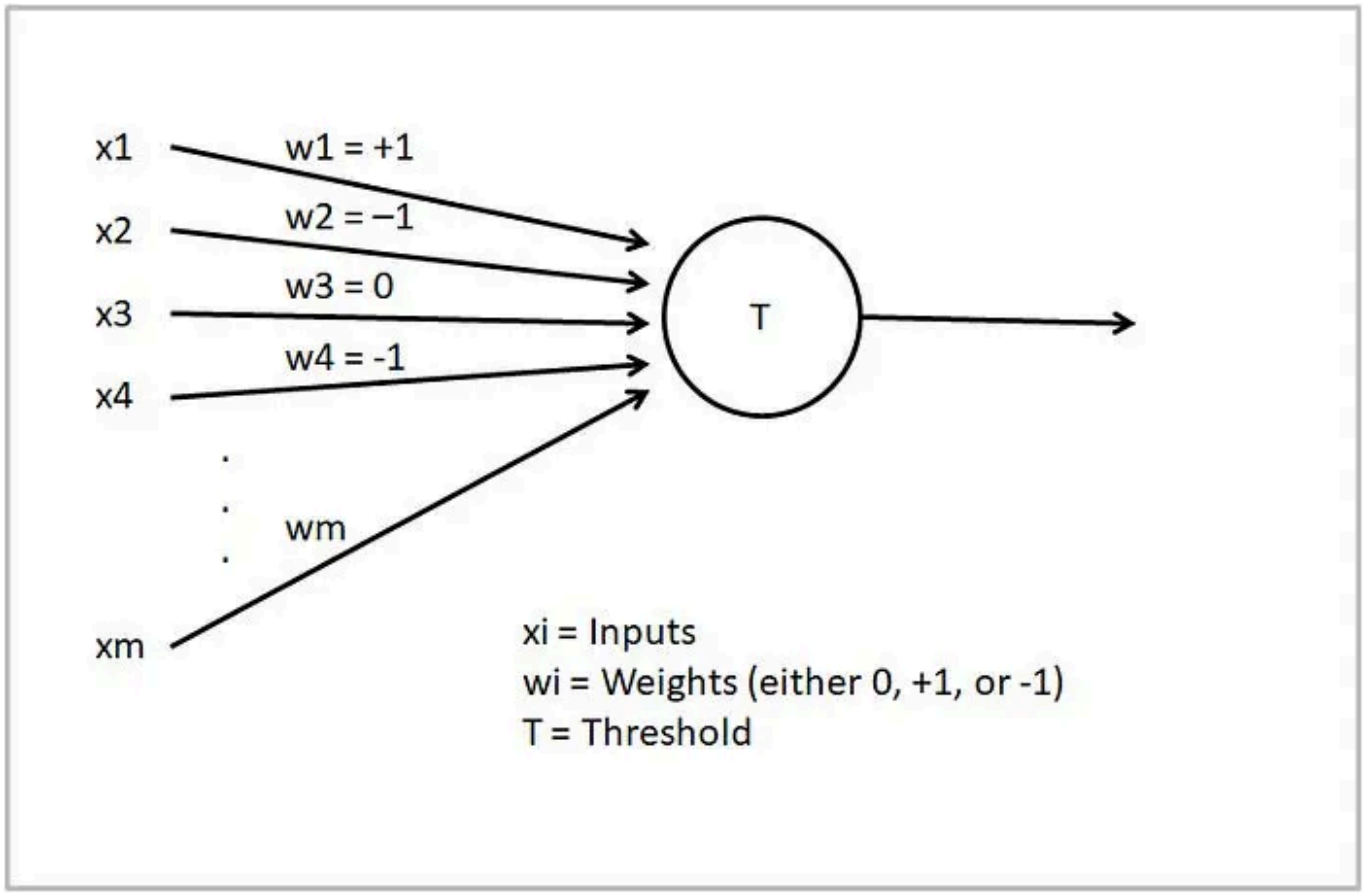
There's something else you need to know that most people don't, which is fundamental to grasping just how deep the rabbit hole goes: Whereas the smaller three subfields — generative AI, LLMs, and transformer-based chatbots, for which ChatGPT serves as a stand-in — were conceived in the last decade (if I were to put an inception date for each, I'd say GenAI corresponds to the "Generative Adversarial Networks" paper in 2014, LLMs to the "Attention is all you need" paper in 2017, and modern chatbots, of course, with ChatGPT's launch in 2022) the other four fields are *much older*. AI was established as a field of research in the summer of 1956. ML and DL were introduced in 1959 and 1986, respectively. And artificial neural networks — a term borrowed from the, at the time, immature field of neuroscience — appeared *the earliest,* in 1943, with the pioneering work of Warren McCulloch and Walter Pitts. Their work is now considered the first ever on artificial intelligence, and that's where our story begins: to heal one's traumas, one needs to go search for the origins.

(AI pioneer Jürgen Schmidhuber would contest every one of these dates, but since I'm no historian, a peer-reviewed genealogy is unnecessary; we only need a shorthand timeline of turning points.)

After all this boring background, I might continue by asking a rhetorical question: What if I told you that a biological neuron is better represented by a whole artificial neural network than by a single artificial neuron? Or, to be less nerdy: Why a field that's concerned with creating intelligent agents that can reason insists on being as stubborn as a rock? Or, to be clearer (last question, I swear): Why has no one revisited the first assumption ever made in AI despite having been proven, time and again, that it's a dangerous simplification of reality?

II.

It was in 1943, during the early days of computer science, that the first neural network was created. Alan Turing, one of the fathers of computer science, had published his foundational paper on computability and the Turing machine just seven years earlier, and it would still take him another seven years to publish his famous paper on the Imitation Game (known as the "Turing Test" which, despite what Gary Marcus says, has clearly been beaten. By 1943, neuroscientists had been modeling biological neurons for decades (at least since the late 19th century, through the work of Santiago Ramón y Cajal and Camillo Golgi), but McCulloch and Pitts' work was the first to describe them in terms of propositional logic; in other words, as a system that could, in principle, be *implemented by a computer*. Those of you familiar with the history of AI will recognize this picture:

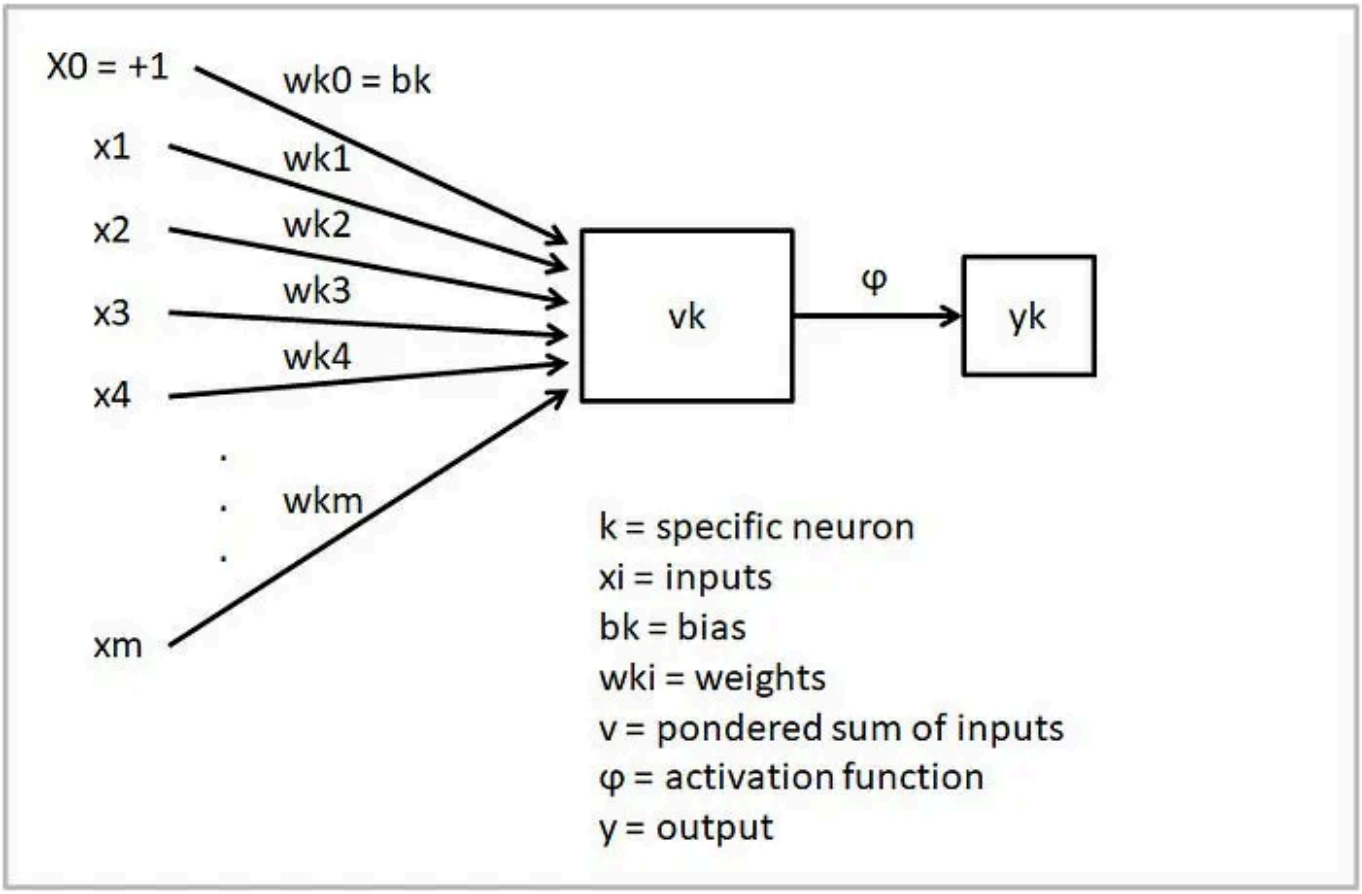


MCP neuron

Simple enough. This is the McCulloch and Pitts neuron model (MCP). It's an 80-year-old fossil. And still, it is the same model that's taught in every modern introduction course on AI. The reason is not so much that MCP is an accurate model that has required no refinement over the years, but that deep learning hasn't changed a bit at the elemental level since 1943. (Scientists love to repeat that "all models are wrong, but some are useful"; the AI field has stretched the implication that "wrong models can work well" to a point that borders on irresponsible, as we will soon see.)

The MCP neuron was intended to represent a simplified neurophysiological version of biological neurons: A series of inputs goes into the neuron, which processes them and then either generates an output or doesn't. It had a threshold activation function: if the sum of the inputs is negative, the output is 0; otherwise, it's 1. Current neural networks (like the transformer

architecture underlying ChatGPT) have weighted inputs and more complex nonlinear activation functions (no need to understand what any of this means) to allow for meaningful learning and more precise input-output mapping. But they're just a slightly improved version of the MCP model. The basis is the same; some inputs are transformed into an output. Whether the neuron learns well or not is irrelevant to the fact that these models don't resemble a real, biological neuron.



Current ANN neuron model

That's why you will hear, not totally inaccurately, that "AI is just linear algebra." At this level, it's true. The problem is over-extending this assertion: Complexities emerge with scale ("more is different" and such), which is why I dislike that kind of criticism; in the sense it's true, it's trivial, but in the sense it might be interesting, it's wrong. It's like saying that Anna Karenina is a book like any other; true, but meaningless. Insofar as the analysis stays at

the neuron level, it's fair to be dismissive of the architecture, but scale creates epistemic voids that no one can bridge. That's why you'll also hear, also not totally inaccurately, that "we don't know how AI works," because modern AI is not designed, coded, or built but rather nurtured and grown, like flowers in a garden or trees in the forest.

But let's get back to MCP. The main simplification — and the one that most hurts the fidelity of the model — is that each neuron is collapsed into a single point in space. This is sufficient to simulate the behavior of some simpler neurons, but the biophysical nature of other, more complex neurons is too nuanced and intricate. In animals, electricity flows through the dendrites, the soma, and to the axons through space and time. Not all dendrites work the same way. Not all inputs participate in generating the outputs (voltage decreases along the dendrites). Dendritic tree morphology, synaptic patterns, and different types of receptors all influence neuronal behavior. And many more elemental mechanisms and processes form the bases that eventually give rise to our intelligence. None of those characteristics is described in the MCP or the neuronal models currently in use. (There have been attempts to modernize AI from the ground up, like neuromorphic research and updated elemental neuronal models — "Expressive Leaky Memory neurons," "Multi-compartment spiking neurons," or "Dendritic spiking neurons" — but although those efforts exist, they gather no funding or interest beyond some obscure corners in academia.)

(You should know that all these weird-sounding neuron names are also *models*. Again, we must invoke the "all models are wrong, some are useful" aphorism. The goal of science is not to reveal the True Nature of the world, which is most likely beyond our grasp, but to reduce the discrepancy between our models and that unfathomable reality while keeping them useful; that's where science and engineering clash. Get too close and you will

have a slightly wrong, but unaffordable or unfeasible model. Stay too far and you will have an excessively wrong model. In case you're curious, the classic gold standard model in neuroscience for mathematically expressing neuronal behavior is the Hodgkin–Huxley model from 1952, which describes the neuron with four coupled differential equations describing ion channel dynamics (Na^+, K^+, leak currents — whatever that means, right?) and is the basis for nearly all later biophysical models that, funny enough, require computers for simulation and whatnot.)

Notably, by the time the MCP neuron was ideated in 1943, neuroscience *already knew* neurons had features that made them non-reducible to a space-point neuron. McCulloch and Pitts simplified those annoying complexities in order to make a logic-based model (a great starting point!). But in the process, they set the groundwork of a whole field that never deigned to look again at its own premises and compare them with relevant discoveries in the cognitive sciences. If you ask why it didn't happen with AI when science — to use the popular paraphrase of Max Planck's principle — "progresses one funeral at a time," meaning that dead ideas are the norm and a requirement for progress, then you'd be asking the right question. And I will answer, not to spoil the later sections, with a parallelism:

It takes a kind of idealistic bravery or foolhardy courage to topple a brittle foundation in academia; it takes a will of steel or the power of an emperor to topple a brittle foundation in industry.

Neuroscience, in contrast to AI's alchemical essence (which is fine), is an established science; being commercially irrelevant, neuroscientists are compelled to follow the scientific method, for they can't follow profits. But the most important contrast is that neuroscience has *developed drastically* in the last 80 years; it was more mature back in the 50s, and it's *much* more

mature today. Our understanding of biological neurons has come to a point where AI researchers can't keep calling artificial neurons "neurons" at all. And indeed, they don't. Everyone working in AI takes for granted that ChatGPT has a distant inspiration in the human brain, but you don't hear them mention the words "neuron" or "unit" anymore, like they used to in pre-ChatGPT times. What began as a useful simplification by the field turned into an intentional omission by the industry. Insofar as LLMs work and ChatGPT is popular, and both provide a return on the investment, it doesn't matter to them.

I will argue, contrary to what intuition suggests, that it actually matters and that the situation we're in — you know, the 1,000,000,000,000 dollars devoted to building AI infrastructure in the form of GPUs designed in Taiwan and datacenters erected in Northern Virginia, and the 1,000,000,000,000,000 data tokens needed to advance the state of the art ever so slightly — is a direct consequence of not having backtracked when the sunk cost was still reasonable.

But before going into that, let's strengthen my case by taking a look at some of the most relevant recent neuroscience research on the topic. It will illustrate, I hope convincingly, just how far AI has diverged from the cognitive sciences and, thus, from the human brain, the only instance of intelligence that we know of.

III.

(Keep in mind that you don't need to follow any technical or jargon-y details here. The important part is that you get a feeling of the sharp contrast between what neuroscience knows about neurons, neural networks, the brain, and intelligence, and what AI knows about it.)

Let's take a quick look at dendrites, the biological neuron's signal receptor. In the early 1980s, Christof Koch and others found that dendritic morphology and synaptic patterns could influence how neurons internally processed the inputs. For a long time, scientists thought dendrites behaved uniformly and passively added the inputs coming from other neurons, but Koch's findings revealed that form and function are much more complex than textbooks suggest.

In 1996, scientists from Columbia University and Bell Labs investigated the role of individual dendrites and discovered that they act as processing units themselves: Dendrites have their own threshold to generate spikes (called dendritic spikes), which is different than the threshold of the whole neuron. That is, neurons aren't simple "logic gates," as the MCP model suggests. Dendrites seem to be capable of acting as logic gates themselves. A biological neuron is therefore a processing system that, in turn, is composed of independent processing systems: processors within processors.

To represent this in artificial neural networks, connections between neurons would need to have distinct morphologies affecting their role in neural outputs (it doesn't happen). Then, those connections would act internally as processing systems (it doesn't happen): Each input connection that arrives at the neuron would generate (or not) a spike that would change the overall output of the neuron (it doesn't happen). This complexity mismatch implies that a biological neuron is better understood as a *layered network*, in which the layers (dendrites) function as nonlinear intermediate input-output mappings. The resulting intermediate outputs are then summed accordingly to the morphology of the "connection tree" to produce the final output.

In 2020, only five years ago, Albert Gidon and colleagues at the Humboldt University of Berlin published a groundbreaking paper in Science, building

on these striking discoveries. They found a new input-output feature in pyramidal human neurons that wasn't described by current models. Dendrites from these neurons produce a type of spike in which intensity is highest for stimuli at the threshold level and lowest when the incoming electrical current is increased.

This discovery proved that some dendrites can act as XOR logical gates; the output is true if and only if one of the inputs is true. In 1969, Marvin Minsky and Seymour A. Papert, AI pioneers, proved that a single-layer perceptron (an early type of artificial neural network) couldn't do this type of computation. Gidon's research proves that *a single biological dendrite can*. An element inside a biological neuron can conduct complex computations that an entire neural network can't. Of course, more complex ones could (we're talking here about dozens of parameters, whereas GPT-4o had hundreds of billions), but this still signals a two-order-of-magnitude computational gap between the basic elements of human brains and AI systems. If a dendrite can do the work of a basic artificial neural network, then how much more complex are biological neurons compared to artificial neurons? Or, a better question, how much more *powerful* are they?

In 2021, a year after Gidon's research, David Beniaguev and colleagues published a paper in Neuron that proved what's been suggested for decades: An artificial neuron can't accurately represent a biological neuron at all.

To prove this, they used modern machine learning techniques to simulate the input-output behavior of a pyramidal human neuron. They wanted to test two things: Whether an artificial neural network can precisely predict neuronal outputs when trained on real input-output pairs, and how big the artificial neural network needs to be to capture the whole complexity of a biological neuron with sufficient accuracy. They found that, at the very least,

a 5-layer 128-unit temporal convolutional network (TCN) is needed to simulate the input-output patterns of a pyramidal neuron at the millisecond resolution (single spike precision). They modified depth and width and found that the best performance was achieved by an 8-layer 256-unit TCN.

To make a gross comparison: This means that a single biological neuron needs between 640 and 2048 artificial neurons to be simulated adequately; that's *three orders of magnitude*.

Notice, because I want to be as rigorous as possible, that this doesn't necessarily imply that a biological neuron has this much more computational power or complexity, but it's a clear sign that both types of neurons are further apart in structure, function, and behavior than previously thought. It is for this reason, and others I won't mention here, that it is a miracle that artificial neural networks, and the entire modern edifice of AI based on deep learning and LLMs, work at all. The "unreasonable effectiveness of deep learning," they call it. The problem with something that is unreasonably effective is that it might break at an equally unreasonable, or rather *unanticipated*, moment.

Beniaguev's team was able to pin down the exact mechanisms by which the biological neuron was so difficult to simulate: dendritic morphology and the presence of a specific type of synapse receptor called NMDA. These are structural and functional aspects of biological neurons well-known in neuroscience for a long time, but have been completely ignored in modern AI and ANNs, except for a handful of studies that seldom leave the classroom. One last example I will share is this paper published in Nature Communications by Panayiota Poirazi's team, who built an artificial neural network with dendritic morphology and connectivity and showed that it presents superior efficiency, resilience, and precision than standard ones.

It's clever and brave in the well-intended, naive way that characterizes academics: they keep trying to bridge the gap between AI and neuroscience without realizing that the bridge was dismantled by the people they're trying to appeal to. This paper, as interesting as it is, attracted a striking 23 citations and none of the funding.

Some questions arise from these insights: Why hasn't the AI community tried to remodel its foundations to better adapt to the reality they're trying to simulate? Is AI destined to fail in its quest to achieve artificial general intelligence (AGI) until those foundations are overthrown and rebuilt from the ground up? What would be the consequences of changing AI at such an elemental level? Or, conversely, what would be the consequences of *not changing anything at all*?

IV.

Let's take the standard position in AI: It may be the case that AI systems like ChatGPT can be scaled and improved iteratively — that's what Google DeepMind, OpenAI, Anthropic, and Meta do — until they eventually reach AGI or human-level performance without any intentional restructuring of the basic assumptions that preceded the current work in generative AI and LLMs (I say intentional because it's always possible to correct superficially while moving forward, just like a dumb, non-teleological evolutionary system would). Perhaps the apparent differences between artificial and biological neural structures are, at worst, reconcilable with enough resources (e.g., compute and data — you know, all those quadrillion GPUs and tokens), and at best, irrelevant.

Neuroscience is concerned with intelligence, the brain, and the mind. Neuroscientists are naturally concerned with looking inwards, to the only instance of intelligence we know of: us. In contrast, AI researchers are

concerned with *replicating intelligence using artificial means*. They care about designing and building intelligent agents that perceive the world and act on it accordingly (I take this as the ultimate goal of the field, even though I know there are intermediate goals that have nothing to do with neuroscience, like Vibes and Sora and the AI slop that's having a feast on our biological neurons, hopefully to replicate them someday). If you look *dispassionately* at this discrepancy, you will realize that both fields are doing the sensible thing. Neuroscience cares about the premise of intelligence, whereas AI, about its outputs. If we can find a different substrate (e.g., silicon vs carbon), architecture (e.g., LLMs vs cortices), or a set of trade-offs that better meet our circumstances, then we needn't model an intelligent agent after the human brain. I'd even go further: it's not possible! Because computers and AI systems already display a series of abilities that humans can't replicate, like near-instant computation, arbitrarily high precision, eidetic memory, etc. In short: humans are one instance in the space of possible intelligences, but in principle, no physical law prevents the existence of others, and, all things considered, I see no reason to prefer ours…

Even people who've grown wary of scale maximalists and LLM maximalists (those are the labels for people who think we already have all the elements we need) but *can't be accused* of hating AI in any way, like scientists Francois Chollet, Andrej Karpathy, and Richard Sutton, to name a few well-known figures in the field, generally agree with this point: There's no need for full biological fidelity if you can achieve functional competence some other way. Here are two tweets by Chollet on his stance on LLMs (2024) and the pursuit of AGI (2025):

> LLMs (and deep learning in general) are useful, and will keep getting more useful over time. They also won't morph into intelligent agents, because they

> simply don't implement the mechanism of intelligence.
>
> The point of our work isn't to build an artificial human. The universe is full of questions far more interesting than our own reflection. The point is to create a new kind of mind to help us explore & understand the universe better than we can ourselves.

LLMs lack key mechanisms for intelligence, but that doesn't mean that they need to resemble an "artificial human" to be intelligent! Here's Karpathy, in a comment on Sutton's interview in the Dwarkesh podcast:

> . . . today's frontier LLM research is not about building animals. It is about summoning ghosts. You can think of ghosts as a fundamentally different kind of point in the space of possible intelligences. They are muddled by humanity. Thoroughly engineered by it. They are these imperfect replicas, a kind of statistical distillation of humanity's documents with some sprinkle on top. They are not platonically bitter lesson pilled, but they are perhaps "practically" bitter lesson pilled, at least compared to a lot of what came before. It seems possibly to me that over time, we can further finetune our ghosts more and more in the direction of animals; That it's not so much a fundamental incompatibility but a matter of initialization in the intelligence space. But it's also quite possible that they diverge even further and end up permanently different, un-animal-like, but still incredibly helpful and properly world-altering.

And here's Sutton on the podcast about how kids don't do imitation but mostly experience things as proactive agents living in the world, whereas LLMs mostly imitate human-made training data:

> It's surprising you can have such a different point of view. When I see kids, I see kids just trying things and waving their hands around and moving their eyes around. There's no imitation for how they move their eyes around or even the sounds they make. They may want to create the same sounds, but the actions, the thing that the infant actually does, there's no targets for that. There are no examples for that. . . . The large language models are learning from training data. It's not learning from experience. It's learning from something that will never be available during its normal life. There's never any training data that says you should do this action in normal life.

They're talking about the same thing in slightly different ways: What we have in the AI field (LLMs) is not enough, and although bioinspiration is worth taking into account, we may end up with something that resembles little "an artificial human" and is more "un-animal-like." It makes no sense to reject a priori any insights that the cognitive sciences could provide (that's Karpathy and Sutton's main argument), but we also don't need to stick to what we already know just because it's what we already know (that's Chollet's point).

To answer the question with which I ended the previous section: Maybe there's no need to revisit the original neuron model from 80 years ago. Maybe we keep going just like we have for the last 10 years, and nothing catastrophic happens. We mustn't make this mistake: AI's goal isn't and has never been *replicating human intelligence* at every level. If a shallow abstraction achieves the goal, why complicate it with dendritic spikes, voltage gradients, and weird morphology? Is complexity a prerequisite for intelligence? That makes no sense; evolution makes things convoluted precisely because it carries no teleological impulse whatsoever, proceeding instead under the modulation of survival pressures and other external forces.

Those research studies I referenced are concerned with how much more complex human neurons are than we thought, but that's not the right lens to look at the problem if you're an AI researcher! Is neuron-level fidelity essential for cognition? I will answer by validating the opposite hypothesis: two instantiations of intelligence of drastically distinct complexity, form, size, substrate, and phenomenology could, in principle, result in comparable intelligence; the map needn't match the territory to produce a usable compass.

These are totally reasonable counterpoints to my starting arguments. After all, McCulloch and Pitts' "good enough" 80-year-old simplistic neural model is the foundational basis of systems smart enough to beat humans at chess and coding, and math, and in an increasingly larger number of problems in diverse fields. Benchmarks that help us evaluate AI models' performance are saturating because we can't come up with tasks that are hard enough. However, there's an important limitation with standard evaluations of AI (I include here those trying to capture the chaotic nature of the real world, like METR's analysis of long tasks and OpenAI's GDPval): AI tests are designed so that *none of the stuff that we're uncertain about is actually measured.*

To be precise, the only relevant benchmark is *competence in the real world,* and so far, every single benchmark fails to capture the complexity of reality and every single AI model fails at this when tested "in the wild" in a way that human toddlers wouldn't (you can see it with benchmarks that require on-the-fly solving skills like ARC-AGI or in large-scale projects that are factored in in productivity and economic analyses but not tested for in laboratory evaluations). There's a huge gap between prediction-competence and environment-competence (Chollet discusses crystallized vs. fluid intelligence, a psychological concept that highlights the distinction between possessing extensive skills and *being adept at efficiently acquiring new ones*).

So maybe, although functional performance is the goal, it turns out that biological fidelity is the requirement to a much deeper degree than we'd like to admit; after all, it's the only reference we have of world-competence.

It's fine to formulate the hypothesis like: "Ok, maybe this plane can fly without wings," even if any onlooker familiar with those little animals called birds thinks it's nonsense. That hypothesis is the equivalent of saying, "Ok, maybe this AI can be intelligent without all those oddly shaped gyri and sulci." But, if after years of attempts and amounts of funding that would dwarf the cost of directly breeding a giant species of bird that could carry us around, you're still unable to make the damn thing fly, maybe it's time to reconsider your premises. Maybe adding wings to the plane is not an unreasonable consideration. But I played a trick on you just now, because I used an analogy to draw the parallelism that *I know* is false. One could also propose this other hypothesis: "Ok, maybe this plane can fly without *flapping its* wings." Oh, wow, that's almost the same thing and yet so different — because it's true! Planes don't need to flap their wings! Having limited knowledge about what works and what doesn't is the root cause of the problem. How much should we take from biology? We don't know because, turns out, our metaphorical AI plane kinda flies and kinda doesn't. And we have no idea why both things are true because the pattern it follows is, itself, completely unfamiliar to us.

So it's pretty clear at this point that the space of possible intelligences is larger than nature allows for (I mean it in the biological sense; in the physical sense no one, not even AI systems, are allowed to disobey, e.g., the speed of light and the second law of thermodynamics are inviolable), and yet we may not want to drift away from the carved path too much. Better safe than sorry, especially if the apology will be accompanied by a trillion dollars in losses. We don't know what elements are always needed and which are

evolution's specific trade-offs that we, intelligent designers, don't need to incur (e.g., planes don't flap their wings because engines create the required propelling force, but still are subjected to the physical constraints imposed by aerodynamics, so the wings stay). That's why playing conservatively in the creation game is the best approach: evolution's design space might contain structural priors that encode principles of intelligence that we take as incidental quirks; you don't play with the unknown unknowns. We're not as free of compromise solutions as we like to think, and we don't know how ours — to some degree similar and to some degree distinct from biology's — affect our arbitrary mapping from design space to solution space.

The map needn't match the territory to produce a usable compass, but the more faithful the representation, the harder it is to get lost. Or, in other words, we don't need to copy nature, but we'd better respect its boundaries (those we deem sensible and those we don't yet understand).

Richard Sutton is best known for his 2019 Bitter Lesson essay, which says, simplifying, that in the long term it's preferable to let computers take the burden of finding better AI systems instead of trying ourselves by leveraging our limited knowledge. And that's fine, probably true. The mistake is having called *that a* bitter lesson. It's bittersweet at most. You know what the true bitter lesson is? That evolution, through non-teleological trial and error and random mutations, managed to scale efficiently a more complex elemental building block into a greater, smarter, more robust, and more versatile system: it expertly carved a seemingly convoluted path from the biological neuron to human intelligence. Sutton's lesson is powerful because it doesn't attend to the complexity of the elements to be scaled, just to the fact that they should be efficiently scalable. Adding compute works better than tweaking weak heuristics and leveraging incomplete knowledge, but nothing beats — and I remark *nothing* — starting from the correct building blocks.

In a way, this lesson reminds me of the main takeaway from *Seeing Like a State*: if you simplify the world too much, you will make it legible, but you will also destroy the very elements that keep it alive. By trying to perfectly control and monitor the artificial brain that you're making through the simplest neuronal model you could find, you have ensured, almost as an axiomatic truth, that it will not be smart in the ways you want. Despite this bitter lesson that both field and industry have ignored for so long, you will keep adding layers to your LLM, so I make this bet: you will run out of layers to add (computing power, training data, etc.) much sooner than you will beat the unnecessary hurdles you incurred by not having revisited the foundations of your field. That I would call bitter.

V.

I've mentioned several times throughout this long piece something that deserves a standalone section, the kind of trade-off that we suffer in every aspect of life, but biology never faced and will never face: the motivations of capital. It's extremely expensive (if possible at all) to model a neuron in its full complexity, so they chose the simplest model they had (that was affordably scalable) and rolled with that. In a way, evolution's trade-off between brain size and brain weight is not that different from AI labs' trade-off between model size and model cost, except for one thing: whereas evolution has no goals and thus no interest in short-term wins (whatever that means in a genetic context), investors have a laser-sharp focus on the prize they're chasing.

So it makes sense to not overcomplicate things early on, but it's put us in a tricky situation: We saved a lot of money by using an extremely simplified model of the neuron, but maybe we're paying for it extra now that it's proven harder to emulate human intelligence in real-world settings.

The hundreds of billions of dollars that unprofitable AI companies are pouring into datacenters to train more and more LLMs are not a testament to their grit, ambition, and vision but to their *haste*. (I won't entertain here the possibility that they don't care about any of this, and they're actually pursuing more prosaic purposes like absolute surveillance; that topic requires a standalone article.) Had they devoted more resources and time in the beginning to finding a better foundation, we wouldn't be wasting so much energy and money in training LLMs on the entire corpus of human data when newborns need an infinitesimal fraction of that to grow into much smarter creatures.

But of course, had AI pioneers done that, no rational investor would have given them a single dollar. If ChatGPT exists, it's because we pay for it; if ChatGPT exists, it's because there was no one reckless enough to aim higher; if ChatGPT exists, it's because the socioeconomic dynamics that govern its existence don't allow for something better. And so we find ourselves, like the very systems we try to transform into truly intelligent entities, trapped in a local minima. (Isn't it ironic that our economic optimizer mirrors the gradient descent that powers AI systems — or perhaps it is a pattern…)

Capital allocation agents do this all the time (if you want to surrender the last ounce of faith in humanity you might still be holding onto at this point, you can read Scott Alexander's "Meditations on Moloch" to finally see the entire rabbit hole at once). Think about the Industrial Revolution, when we decided that automation was worth polluting the world. (I will also not turn this into a climate change debate.) Back then, we didn't have a real-world model on which to base our progress. How could we have known that burning dirty coal would be bad for the lungs of our little green planet, and instead should have invested in nuclear, solar, and wind power early on? Besides, we shouldn't forget that innovation has other limitations; it can only push

frontiers into the adjacent possible thing, not further. But the AI field is special because it has access to the best inspiration it could ever need (I mean us, humans, although I'd understand if you considered this a controversial affirmation), and yet capital allocation agents are *still making the same mistake*: they allow optimization and financial pressures to steer us away from the safer, more sensible path.

Unfortunately, science follows what's fundable, measurable, and publishable, and that's a hard trade-off to navigate because from the short-termist bird's eye view of a capital allocator, what looks like the better approach may prove to be, for reasons beyond its gaze, the most expensive one; pushing forward those trade-offs it doesn't want to take until it's forced to take them — the very essence of capitalistic dynamics — is what brough us from there to here.

What do I mean by "from there to here"? If you've been paying attention, I've poured 7,000 words and then some into drawing an association between the 1943 pioneering work of McCulloch and Pitts and the 2025 financial catastrophe that's looming because AI companies have gathered so much opportunity and so much sway that they can do whatever they want, which is not what they should. And they happen to want to spend hundreds of billions of dollars in building AI infrastructure that only needs to be so expensive because what appeared, 80 years ago, like a simplified model of the computational unit — and thus a welcome news in the eyes of capital — turned out to be an expensive mistake when the industry realized that it needed all the data humans have ever produced and the electricity bill of a mid-sized country to train and serve one of those LLMs that's smaller and dumber than a human brain, which would consume, other things being equal, the energy of a home bulb.

To quote another of LLMs' critics and computer science giant, Grady Booch, responding (in a now deleted tweet) to OpenAI CEO Sam Altman about the $7 trillion he said he'd need to "reshape the global semiconductor industry": "If you need $7 trillion to build the chips and the energy demand equivalent of the consumption of the United Kingdom, then — with a high level of confidence — I can assure you that you have the wrong architecture."

The rewards of practical, empirical success outweigh the epistemic cost of theoretical impurity until they don't. When they say that the invisible hand of the markets is an existing example of docile superintelligence, you can use this example as a counterargument. And, if you're still convinced that we should build an actual godlike superintelligence, whatever that is, I hope that I have convinced you that we may want to rethink the foundations we're standing on a little bit more than we'd otherwise do. I don't think it's a reasonable goal for the McCulloch and Pitts neuron to become God.

Contrary to what it might seem from this essay, I don't resent the choices the AI field made. It's fine that they didn't pay attention to biology's trade-offs, because the human world has its own. But wouldn't it be darkly ironic if, in ignoring nature's warnings — she's wiser than we'll ever be — we ended up digging our grave? It's not clear whether one can draw a reasonably precise trajectory between these two realities — that McCulloch and Pitts' neural model was too simple, and that we're now facing a financial disaster because AI won't deliver — but it is striking that, despite all the years that have passed, it's still trivial to write a story like this and argue, with some firmness that the shadow of that first metaphor we failed to do justice to is still chasing us.

Karpathy is right when he says that LLMs are ghosts we have summoned here from another dimension. For your farthest mistakes are the ones that

haunt you the longest.

*Join 40,000 others in **The Algorithmic Bridge**, a blog about AI for the people.*

Artificial Intelligence | AI | Technology | Tech | Science

Written by Alberto Romero

48K followers · 137 following

AI & Tech | Weekly AI Newsletter: https://thealgorithmicbridge.substack.com/ |
Contact: alber.romgar at gmail dot com

Follow

Responses (6)

Bgerby

What are your thoughts?

Eddy Borremans he

1 day ago

Delightful article!

"Is neuron-level fidelity essential for cognition?" -> The billion dollar question! Or at least "Is MCP level fidelity sufficient for cognition?". Probably not, and more probably not in a sustainable way.

I'd argue we have to keep... more

8 Reply

Eddy Borremans he
1 day ago

> It's surprising you can have such a different point of view. When I see kids, I see kids just trying things and waving their hands around and moving their eyes around. There's no imitat...

Fwiw, I wonder if the distinction Sutton makes between imitation and exploration overstates the difference. Children's exploratory behavior, while spontaneous, is still shaped by evolutionary priors. In a sense, it's a form of structured randomness... more

1 Reply

John Thomas Seyman
2 hours ago

Very informative and readable.

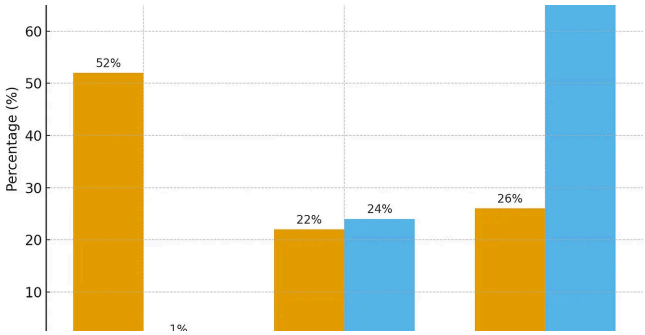
As a trog, I have to wonder, however, whether the rudimentary and inadequate artificial model of the human neuron might nevertheless be sufficient to create AI that will put half of humanity (or more) out of a job.

Damn the logic gates! full speed ahead!

Reply

See all responses

More from Alberto Romero



Alberto Romero

OpenAI Researchers Have Discovered Why Language Model...

A review of OpenAI's latest research paper

Sep 19 908 38

Alberto Romero

The Trillion-Dollar AI Bet

There are bubbles of excess and bubbles of pure betting; some bubbles are both

Oct 2 492 11

Alberto Romero

How to Live Without Your Phone

If you don't do it for yourself, do it for your children

Oct 2 556 16

Alberto Romero

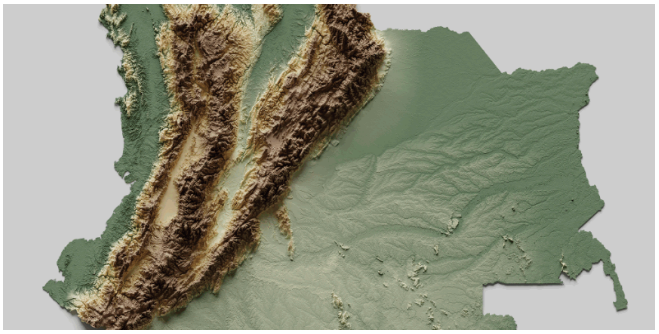
Wherever I Go ChatGPT Follows Me

The internet is not dead, we're just lynching it

Sep 19 242 2

See all from Alberto Romero

Recommended from Medium



Tomas Pueyo

In The Generator by Jim the AI Whisperer

Why Warm Countries Are Poorer

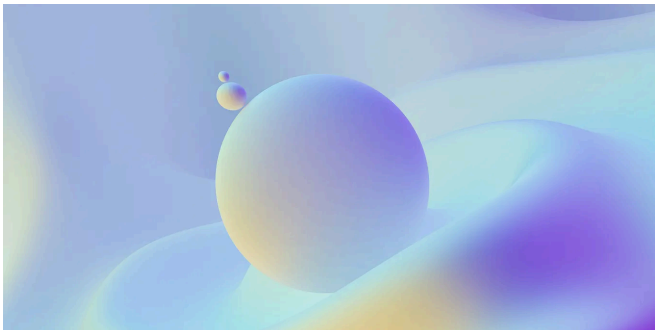
The most underrated factor

Sep 29 4.2K 160

The words “blah blah blah” increase AI accuracy

Who needs Chain of Thought when “blah blah blah” works?

5d ago 3.5K 73

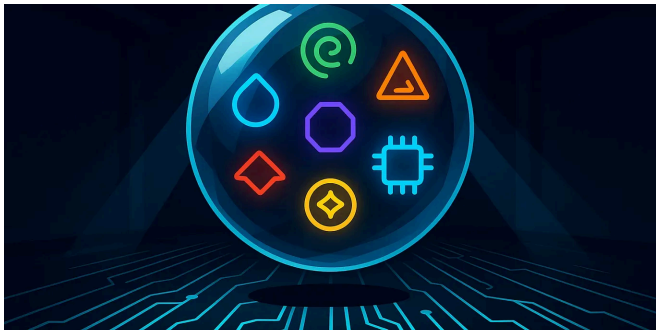


Will Lockett

Is The AI Bubble About To Pop?

The $100 billion red flag.

3d ago 1.8K 55



In Predict by DefineCode

The Future No One Wants to Admit: 7 Predictions About AI That Will...

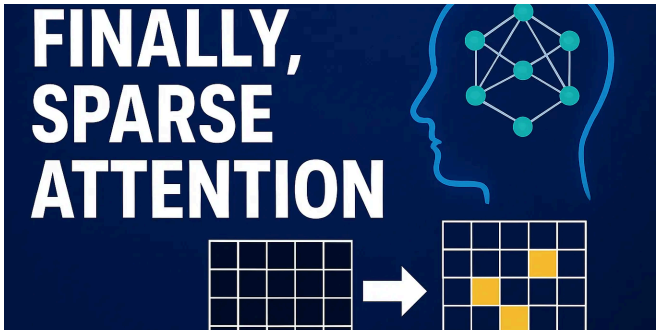
Let’s rip the band-aid off. Most people talk about AI like it’s a shiny toy. Chatbots, art...

6d ago 432 12

B

The End of Degrowth

...or what comes after the realization hits that we’ve been in degrowth for fifty years now



Ignacio de Gregorio

DeepSeek is Finally Back, Solving Sparse Attention.

A Years-old Mystery, Solved

4d ago 550 13 2d ago 780 14

See more recommendations