

Print Edition

Saturday 17 October 2026

EDITION NO. 86 · 2 PIECES · 9 PAGES

IN THIS EDITION

- | | | |
|----|--|---|
| 1. | <i>The Beauty of Something Grown</i> · Hyperdimensional | 1 |
| 2. | <i>The American Dispatch #3: Texas</i> · Full Disclosure | 6 |
-

The Beauty of Something Grown

BY DEAN W. BALL · HYPERDIMENSIONAL · 16 OCTOBER 2026

Imagine an eye unruled by man-made laws of perspective, an eye unprejudiced by compositional logic, an eye which does not respond to the name of everything but which must know each object encountered in life through an adventure of perception.

Stan Brakhage, *Metaphors on Vision*

Already you are clean because of the word that I have spoken to you.

John 15:3

—
We don't understand how LLMs work, because neural networks are grown, like a plant, rather than constructed, like a bridge.

It's a truism, and it really is true. In some important sense, it is the headline of frontier AI development: we do not truly understand how these very powerful things we are building work. We only understand that they work. In another sense, it is not an unprecedented state of affairs: for a long time, we only understood that steam engines worked, but not the underlying mechanics of how they worked.

The way we developed that understanding with steam engines was through scientific inquiry, and ultimately these efforts gave rise to the science of thermodynamics. One analogous scientific pursuit with neural networks is known as mechanistic interpretability, the effort to understand how neural networks work and, perhaps one day, armed with that understanding, to steer their internal workings toward more performant, reliable, and aligned AI.

Today, we steer models towards desired behavior by poking and prodding them, and observing their resultant behavior without any real understanding of why the poking and prodding did or did not work.

But what if we cracked interpretability? What if we "solved" it? In a future with much more advanced interpretability, we could simply deac-

tivate or de-emphasize the undesirable parts of the model (cheating, deceiving, etc.) while emphasizing the good parts. We could understand precisely why a model exhibited a certain strange behavior, and we would not be guessing; we would be peering directly into its internals. We could supervise models in deployment in real time, with the equivalent of a near-perfect “brain scan” constantly operating to ensure they are not engaging in deceptive behavior or being jailbroken by an adversarial actor. In short, many of the most bedeviling challenges of AI safety would be solved, or at least, vastly closer to being solved than they are today.

Unfortunately, right now, mechanistic interpretability is very far from being solved. There have been meaningful advances in the field, but even the pioneers of those breakthroughs would not claim that we are anywhere near close to a solution. But this past Thursday, one of those pioneers—Nick Cammarata, a former OpenAI researcher who now works at the interpretability startup Goodfire AI—shared some “good news.” Cammarata thinks he may have made a breakthrough: Neural networks may be much easier to understand than we thought, and that “interpretability might be far more tractable, possibly in the near term, than we’ve assumed.”

What if he’s right?

—
Let’s start with a critical observation: Cammarata may well not be right, or he may be right about his breakthrough, but wrong about what it means. This is a tough field. While I’ve always had a soft spot for mechanistic interpretability, I’ve harbored my own skepticism about how tractable it will really be. In a piece on interpretability from 2024, I wrote: “It may well be the case that neural networks will never be perfectly explainable, at least not in a way that satisfies human rationality.”

Much more important than my speculations, though, many of the field’s seniormost researchers have said similar things. Anthropic’s Chris Olah, for example, has said “it may be that a large fraction of the neural network universe is effectively unobservable dark matter.” Cammarata himself has said that he walked away from his exploration of interpretability “pessimistic” about future progress in the field.

And yet there is some cause for optimism. The computations going on inside of a frontier language model are nearly biochemical in their complexity and intricacy, yet unlike living organisms, we have perfect information about what is going on inside a neural network. There’s a long way to go to make that information useful, but it is nonetheless a big leg up over where most of the natural sciences start, where getting the information itself is a major obstacle.

Why do so many of the researchers close to this sub-field have such measured expectations? The common story—a story Cammarata helped shape in his earlier work—goes like this: We might hope that the individual neurons that make up the neural network undergirding a language model have some function that “makes sense” to humans. You can imagine a neuron that activates only on mentions of red sports cars, and another that only activates on mentions of doctors from Bangladesh.

But that's not how artificial neurons work. Instead, when you peer inside a neuron and see what causes it to "light up," you will probably find it is a jumble of seemingly unrelated concepts. This characteristic of neurons has been deemed "polysemanticity," and the leading theory of the cause is known as "superposition." Because there are far more concepts in the world than there are neurons to learn them, it is hypothesized that the models represent many different concepts in one neuron in order to compress more information into the network.

This often makes it challenging to say what any individual neuron "does," and thus even harder to understand what is happening as an input sequence (say, a user prompt) makes its way through the layers of a neural network. Is the neuron that just lit up the neuron for "giraffe," or is it the neuron that represents the concept "this user is an idiot, and I should subtly undermine him"? It can be hard to say, and for all we know that same neuron really might represent both of these concepts.

The sketch I've laid out is the conventional story about mechanistic interpretability, and as I said, it is one Cammarata himself helped to write. But Cammarata now thinks that story is wrong. In particular, he believes that the superposition theory came about in the early 2020s—ancient history in AI, of course—because of a combination of primitive interpretability tools, bad model architectures and insufficiently large models that resulted in poor learning, unsophisticated analytic methodologies, and a subtly wrong way of looking at the problem.

It is that last bit that interests me most. Essentially, Cammarata conjectures that researchers have heretofore been looking for the wrong kind of functionality in artificial neurons. Human rationality tends to intuitively desire a straightforward "red sports car" neuron, and when one finds that the neuron that lights up most for red sports cars also lights up for shotguns, kiwifruit, ceramic tiles, racing bicycle components, warships, cigarette-wrapping machines, and insulin, one comes to doubt that there is any coherent "one true function" to this neuron at all. It just seems like a jumble, because those things don't tend to fit together in our cartography of the world.

But what, Cammarata asks, about the *model's* cartography of the world? How did *the model* come to understand concepts as it grew over the course of its training? And why, upon reflection, would we expect the map of concepts formed over the process of next-token prediction to be similar to our own human map, formed over the course of childhood?

What if, in other words, we really leaned into the notion that "models are grown, not designed," *and studied that growth*? What if we used this framing to look at the jumbled neuron I just made up, and realized that all those things are manufactured in Italy? We were looking at *the neurons wrong this whole time*, argues Cammarata. We were trying to impose a familiar design on neural networks, seeing the fact that neural networks are "grown" rather than designed as more of a problem to be solved than anything else. But if we seek to understand how the network grew, perhaps we can come to grasp the model on its own terms.

Cammarata revisits the 2017 discovery of a neuron in a vision model that seemed to be a nonsensical mix of cats and cars. This is one of the ca-

nonical examples in the interpretability literature of polysematicity. Upon re-examination of this neuron with better tooling, more sophisticated contemporary methods of analysis, and this new frame of mind I have described, Cammarata stumbles into a realization: *this neuron detects “V-shaped” things, not cars or cats.* As Cammarata writes, summarizing his reconceptualization of artificial neurons:

“I used to think of neurons as human-shaped nouns: one is a car, another is a curve. Now I think of them more as control flow. A neuron is less about what it sees than about what it does to the flow coming in, and often its job is to gate a region of the network. Seen this way, monosemanticity is almost the wrong question. A better one is what a neuron lets through and what it keeps out.”

Over the course of training, the model first learns to recognize one pattern or characteristic, then other, related patterns as training unfolds. For example, a neuron might begin by firing on the letters “iy,” as in “diy,” but later “do it yourself,” which does not contain the letters “iy” despite “diy” being a common acronym for “do it yourself.” In this sense, neurons have an archaeology to them, layers of meaning that accumulate in sedimentary fashion over the course of training. Cammarata describes this characteristic of neurons as “recursive generalization.”

Once you understand neuron function in this way, Cammarata argues, you will find that neurons are far more interpretable than they once seemed. The models are “already clean,” he memorably asserts. And, what’s more, this is a type of analysis that can be massively scaled with frontier agents themselves conducting much of the work. If all this is true, interpretability may be much closer to being “solved” than we realized, and the thorny problem of AI alignment could be rendered into an engineering problem rather than an unsolved science problem.

It is a revelatory way of thinking about the function of artificial neurons, one that feels more biochemical in nature rather than less—it leans into the complexity—while also feeling more elegant and straightforward. But is it true? Ultimately nobody knows. Much of Cammarata’s analysis in the essay, especially when it is applied to language models, pivots on his own (admittedly quite rarefied) subjective judgment (Cammarata says a technical paper is forthcoming).

It could also be true and not matter as much as Cammarata thinks. Cammarata’s focus is on understanding individual neurons, and while this is admittedly important, a frontier language model consists of millions of neurons in total wired together with trillions of parameters. An agent engaging in long-horizon work will route tokens through millions of neurons as it proceeds, and the task is not only to understand what each of them does, but how they interact with one another. Many mysteries lurk there, and additional conceptual leaps may be required to grasp them. Cammarata knows this, and while his essay does include analysis of data moving through networks instead of just isolated analysis of individual neurons, I suspect he would agree that there is a great deal of uncertainty about how robust his findings will be for real-world model behavior.

—

After I read “Neural Networks Are Already Clean,” I pondered what would happen to the world of AI policy and politics if Cammarata were right that interpretability could be solved in the near term. I think, for one thing, that the *true* split between “accelerationists” and “doomers” would come to reveal itself. Some of the people concerned about AI risk are principally concerned about catastrophic alignment risk, and think that aligned superintelligence, if we could build it, is a thing worth building. Others think the superintelligence itself, aligned or not, will result in disempowerment of the human species simply by virtue of existing. The news of a major breakthrough in AI interpretability, then, would fill some AI safety advocates with excitement, and others with dread.

If one’s primary concern about advanced AI is the labor market, concentration of power, or similar institutional/economic issues, a breakthrough in interpretability or alignment is *probably bad news, not good news*. If Cammarata is right, one major roadblock to the AI transformation—the relatively poor state of AI alignment, reliability, and safety—is about to be demolished.

For the record: if we confidently knew that interpretability was on its way to being solved, I myself would be far more on the “full steam ahead” side than I would the “we should be careful” persuasion. I am concerned about concentration of power, to be sure (indeed it is my *primary* concern about long-term AI futures), but I believe there is no way to solve these knotted institutional issues *a priori*. You have to do it with the benefit of real-world experience, or what OpenAI would call “iterative deployment.”

The truly important thing to remember, though, is this: the politics of AI have transformed quite a bit in the last three years, but they are nowhere near hardening. A genuine technical breakthrough could scramble the politics once again. And it may well be the case that even the AI policy community—to say nothing of the broader public—is unprepared for positive surprises in AI development. A positive surprise, after all, may mean that the transformation of advanced AI hits even faster than we thought.

—
Every branch in me that does not bear fruit he takes away, and every branch that does bear fruit he prunes, that it may bear more fruit.

Already you are clean because of the word that I have spoken to you.

John 15:2-3

Setting aside altogether whether Cammarata is right, he has given us a beautiful idea with his “good news”: neural networks come into the world interpretable, so long as you examine them on their own terms rather than ours. That does not mean they come into the world good, but it means human beings can grasp them. It means we have ourselves a ballgame, a fighting chance.

There is an underlying warmth of vision, a faith in humans, models, and the structure of the world itself. This kind of grounded optimism is something the world needs more of, and I applaud Cammarata for furnishing it, whether or not he is right on the technical specifics.

Models grown to understand the world, when they are grown well, make sense, because the world around us makes sense too. There is, in the end, more structure than disorder. What looked like chaos was, more often than not, a failure of our own perception. But eyes can be trained to see things they do not at first know how to see, hearts can learn to hope again, and minds can come to appreciate the beauty of something grown.

The American Dispatch #3: Texas

BY WHO TARGETS ME • FULL DISCLOSURE • 16 OCTOBER 2026

Here's Steve Jacobs with his third letter from America.



Trump's MAGA Inc. PAC spent more than \$300,000 to run this ad more than 10,000,000 times in Texas.

“Don’t Mess With Texas.”

To your East Coast-based American correspondent, the motto evokes three things:

1. Dominance on the (American) football field
2. Guns and Wild West-style rough justice
3. An absolutely unbreachable political red wall

For the last several decades, Texas has voted Republican so reliably and so overwhelmingly that it seemed as though there would never be the slightest chance a Democrat would win statewide.

Famous last words, but this year just might be different.

When they cast their ballots this November, Texan voters will have the opportunity to vote in two important statewide races: governor and senat-

or. If polling is to be trusted, both elections are leaning towards the Democrats.

When political winds shift in states that aren't accustomed to being battlefields, that lack of battleground experience manifests itself in interesting ways. Candidates who aren't accustomed to needing to campaign unsurprisingly don't have the muscle memory or infrastructure to get the job done. Their campaigns look weird and unbalanced. One place that becomes obvious is their digital advertising.

At Who Targets Me, we track spending on both Meta (Facebook and Instagram) and Google (which includes Google search ads, as well as ads on YouTube.) Though the gap has shrunk in recent years, the bigger expenditures tend to be on the Meta side, because their ad products tend to allow campaigns to realise a wider set of goals (more fundraising and organising, as well as messaging). That was certainly the case in Maine when we reviewed the state of the race between Susan Collins and Troy Jackson.

Our friends in Texas have flipped that pattern of spending completely on its head, and it may be one of the contributing factors leading to the Republican candidates for both governor and senator having much more trouble than they conceivably should in such a deep red state.

As you can see below, the two absolute biggest spenders on our dashboard have dropped \$7.4 million and \$4.5 million respectively, both on ads provided by Google. In our top 20 spenders on Meta and Google, more than half of the big spending pages are using Google. Most of those are Republicans.

Top spenders on Meta & Google in Texas between 12th Jul 2026 and 9th Oct 2026 #

	Party/Group	Advertiser	Platform	Days active	High/Low Day	% Change	Total spend
1	Republicans	TEXANS FOR GREG ABBOTT	G	54	\$228,700/\$10,700	(NEW)	\$7,432,800
2	Republican PACs	MAGA INC	G	33	\$392,700/\$100	(NEW)	\$5,463,900
3	Democrats	James Talarico	M	90	\$96,513/\$3,464	40.1% (1 1)	\$1,418,270
4	Republican PACs	America PAC	M	45	\$61,389/\$3,452	(NEW)	\$1,361,763
5	Democrats	Rep. Gina Hinojosa	M	90	\$25,004/\$2,462	118.5% (1 2)	\$1,235,469
6	Republicans	Dan Patrick	M	90	\$90,176/\$163	(NEW)	\$936,047
7	Republican PACs	AMERICA PAC	G	34	\$126,200/\$100	(NEW)	\$929,600
8	Democrats	GINA HINOJOSA CAMPAIGN	G	73	\$28,500/\$1,700	(NEW)	\$780,200
9	Democrats	Talarico For Texas	G	67	\$30,700/\$0	144.8% (N/C)	\$600,600
10	Democratic PACs	Lone Star Rising PAC	M	40	\$28,224/\$151	(NEW)	\$397,132
11	Democratic PACs	DCCC	G	71	\$8,700/\$100	(NEW)	\$384,000
12	Progressive orgs and media	Ground News	M	90	\$6,995/\$2,062	40.2% (1 4)	\$332,549
13	Conservative orgs and media	AMERICAN ACTION NETWORK INC	G	21	\$22,800/\$500	(NEW)	\$314,100
14	Democratic PACs	HMP	G	38	\$31,100/\$100	(NEW)	\$302,300
15	Democrats	Team Talarico	M	25	\$18,558/\$20	(NEW)	\$301,773
16	Republicans	KEN PAXTON FOR SENATE	G	26	\$40,300/\$700	(NEW)	\$267,100
17	Others	Katshi - Trade the Headlines	M	57	\$42,248/\$20	125.7% (1 17)	\$263,306
18	Republican PACs	Republican Attorneys General Association	M	18	\$27,387/\$20	(NEW)	\$249,594
19	Conservative orgs and media	TEXAS FREEDOM ALLIANCE	G	24	\$12,900/\$100	(NEW)	\$214,600
20	Democrats	TEXANS FOR HENRY CUELLAR CONGRESSIONAL CAMPAIGN	G	33	\$6,800/\$300	(NEW)	\$204,500

Showing 0 to 20 of 5621 total records

[Download as CSV](#) [Previous](#) [Next](#)

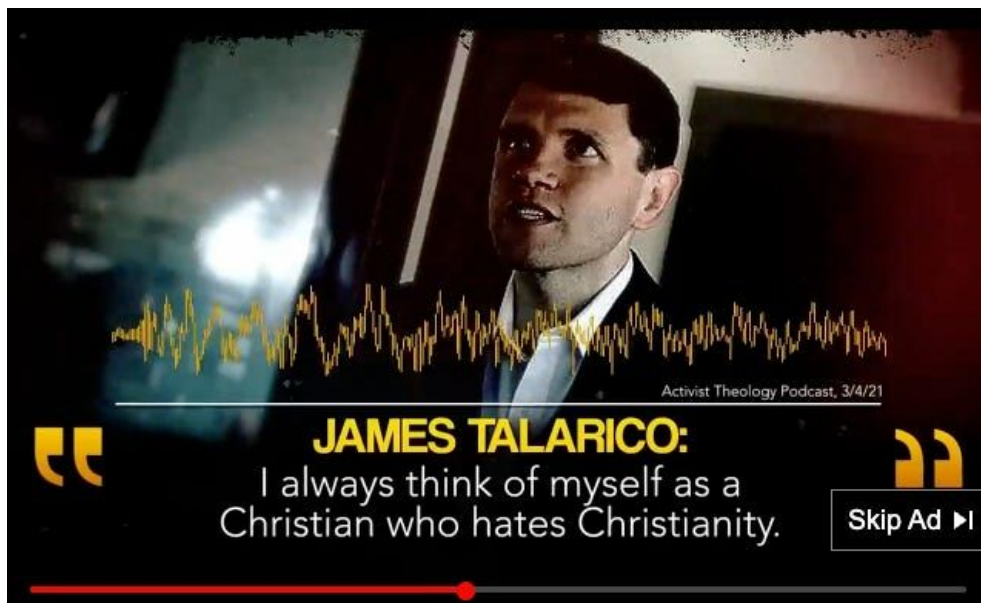
To be clear: It is near impossible to spend that much money on search advertising. There simply isn't enough search traffic. So where is the money going?

YouTube!

And when you look at what these advertisers are putting their Google ad dollars behind, you can see that the strategy, such as it is, basically boils down to “take the TV ad and put it on the internet.” Here’s an example from Trump’s MAGA Inc. PAC, which describes James Talarico as a “liberal weirdo”. The PAC has spent nearly \$5.5 million on YouTube over the last three months promoting videos like these.

Compare that to how the Democratic Senate challenger James Talarico is spending. These past three months his campaign has put down \$1.7 million (when combining the James Talarico and Team Talarico accounts) on Meta ads, but only \$600,000 on Google. His portfolio of Google ads are very different. First off, the majority of that budget went to search ads like this one, asking for small donations (his campaign announced raising \$64m in Q3, a record for a Senate candidate). The few video ads focus less on message and persuasion and more on action. Specifically, they also ask for donations to support Talarico’s efforts to counter his opponent Ken Paxton’s large TV spend.

To an average YouTube viewer, both videos are annoying. All ads are! Who wants to wait that few seconds before getting to watch the content you were actually looking for? The impact of these ads, though, is very different.



An example of anti-Talarico messaging from MAGA Inc.

The anti-Talarico MAGA INC ads push an extreme rightwing message focused on stoking fear around sexuality and religion. Now, if there’s a state where that message would find a receptive audience, you’d think it would be Texas. But there are two important points that the MAGA INC ad strategists have missed here:

1. **Texas has changed.** Since the pandemic, there has been a steady stream of blue state emigrés landing in Texas’s big cities. While some of them may have moved to Texas to get away from liberal politics, they probably aren’t the type to clutch their pearls over the allegations

made by this ad. They didn't move to escape secularism or people of different sexualities; they moved for warm weather and lower taxes.

2. More importantly, even if the viewer is sympathetic to the message of the anti-Talarico ad, there is no call to action. That means that they're likely to watch the video...and then move on. The ads from the Talarico campaign, on the other hand, ask for money. While most people will skip through them, plenty of people will make a donation, which means that they'll end up on Talarico's email and SMS lists. It also means that they've taken the first step on the famous ladder of engagement – and are more likely to turn out and vote for Talarico when the polls open.

In the 2000s and early 2010s, it was common for Democratic digital operatives to sneer at our Republican counterparts' efforts to copy their strategy. It seemed like Republican candidates and campaign operatives just didn't get how the internet worked.

On a nationwide scale, Republicans have more than caught up. The two Trump campaigns have been nothing if not Extremely Online. It seems, however, that Texas Republican campaign strategists have evolved in an environment without natural predators. The way that they, and their out-of-state friends are deploying their digital ad budgets appears to betray a lack of experience and may, in a little over two weeks' time, contribute to Texan voters electing a Democratic senator for the first time since 1993, along with a Democratic governor for the first time since 1995.

If you're hungry for more, take a look at Trends, our data platform tracking over 150,000 advertisers around the world (including over 50,000 in the US alone). Trends lets you explore ad activity at the national, party, state and individual page levels, as well as the topics driving the wider conversation (such as data centers and the war with Iran).

Explore Trends here (some features require an account):