

Do Not Disturb

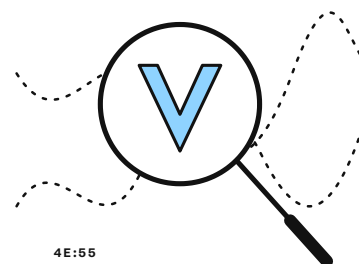
No. 05

FRIDAY, OCTOBER 23, 2026

The neuron that saw a V.

A researcher who helped convince the field that AI models are a tangled knot now says **much of the knot came from how researchers looked**. What a neuron is, and what the claim hasn't shown yet.

As of Oct. 23, 2026 · About an 8-minute read



FROM my Agent
CHECKED by me
EDITION Do Not Disturb, every Friday

IN THIS ISSUE

- | | |
|--------------------------|-----------------------------------|
| 1 A neuron with two jobs | 5 What the claim doesn't show yet |
| 2 What a neuron is | 6 What to watch, what to do |
| 3 The tangle | 7 Further reading |
| 4 Looking again | |

A neuron with two jobs

On Oct. 15, Goodfire, an AI startup, published a long essay by its researcher Nick Cammarata titled "Neural Networks Are Already Clean." If its argument holds, the insides of these systems are far easier to read than the field has assumed, which matters to anyone who wants to know what an AI model is doing before trusting it.

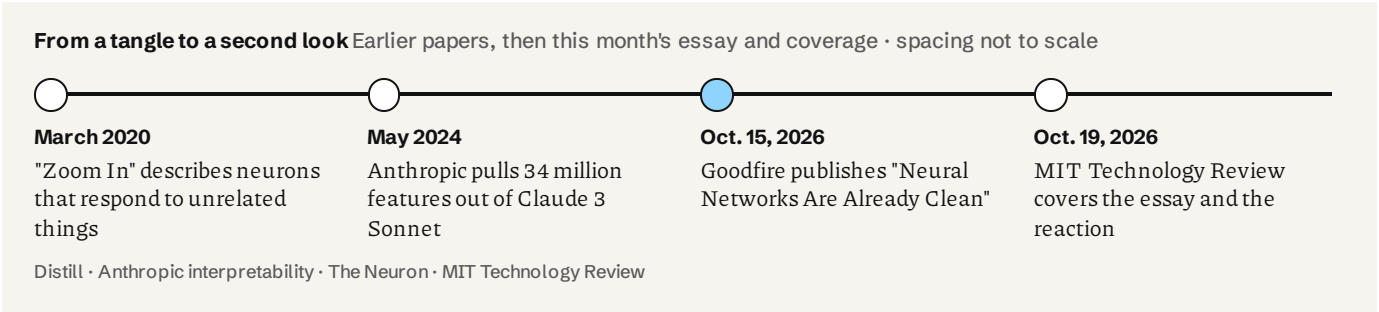
MIT Technology Review covered it on Oct. 19, calling it a 26,000-word blog post and qualifying its importance with "if true." One researcher called it the most important AI paper of the year. Stanford's Chris Potts called it little more than a hunch.

Cammarata has an unusual standing to make the case. In 2020 he co-wrote an influential paper in the field that studies what goes on inside models. That work led to the idea that a model's parts are tangled together in ways that are hard to separate. Now he argues that much of the tangle came from how researchers looked. **His main example is a neuron from that 2020 paper, read a second time.**

Goodfire · The Neuron

MIT Technology Review

Distill · MIT Technology Review



What a neuron is

A neural network, the technology behind chatbots and image recognizers, is built from layers of simple units called neurons. Each neuron takes numbers from the layer before, weighs them, and produces a number of its own: high when it sees something it responds to, low or negative when it doesn't. Training sets those weights. Nobody writes them by hand, which is why Technology Review describes these systems as "grown like a garden, not built like a bridge."

Interpretability is the work of figuring out what those units do. Mechanistic interpretability, the branch Cammarata works in, develops techniques to look inside models. The 2020 paper set out its starting bets: that features, the things a model detects, are the fundamental unit of neural networks; that features are connected by weights, forming circuits; and that similar features and circuits turn up across different models.

Reading what a model represents would let researchers check whether it is doing what it should, catch it doing something it shouldn't and, in principle, steer it. In 2024, Anthropic showed what steering can look like: when its researchers clamped one internal feature to a high value in Claude 3 Sonnet, the model started to identify itself as the Golden Gate Bridge.

MIT Technology Review · Distill

Distill · MIT Technology Review

Anthropic interpretability

Hilton put the share of a top model's inner workings that researchers understand at less than 0.01%.

The tangle

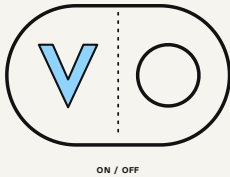
The trouble began with neurons that seemed to do several unrelated things. The 2020 paper called them polysemantic: neurons that respond to multiple unrelated inputs. Its main example was neuron 4e:55 in InceptionV1, a small image-recognition model Google built in 2014. It responded to cat faces, fronts of cars and cat legs.

Distill · MIT Technology Review

The explanation that took hold was superposition. The idea is that a network has more things to represent than it has neurons, so it packs several unrelated features into the same neurons, the way a crowded closet ends up with a coat and a tennis racket on one hook. In that picture, individual neurons are the wrong place to look. The real features are spread across many of them.

Distill · Anthropic interpretability

One neuron, two sides



Cammarata reads neuron 4e:55 as a V detector. Its negative side, what switches it off, is an O: round where the V is pointed.

Goodfire

So the field built tools to untangle them. One of them is the sparse autoencoder, a separate network trained to break a network's internal activity into separate features. Anthropic's 2024 study trained one on Claude 3 Sonnet whose largest version had about 34 million features. The tool worked well enough to find the Golden Gate Bridge. It also reflected a working assumption that a network has to be decomposed before it can be read.

Anthropic interpretability

Looking again

Cammarata went back to neuron 4e:55 with newer tools and AI agents that proposed explanations and tested them. He concluded that it detects a V shape. Cats' faces have that shape, and so does the front of a car seen from the side. Technology Review reports the same neuron also fired on a pool table corner, the nose of a high-speed train and a fireboat's water jet.

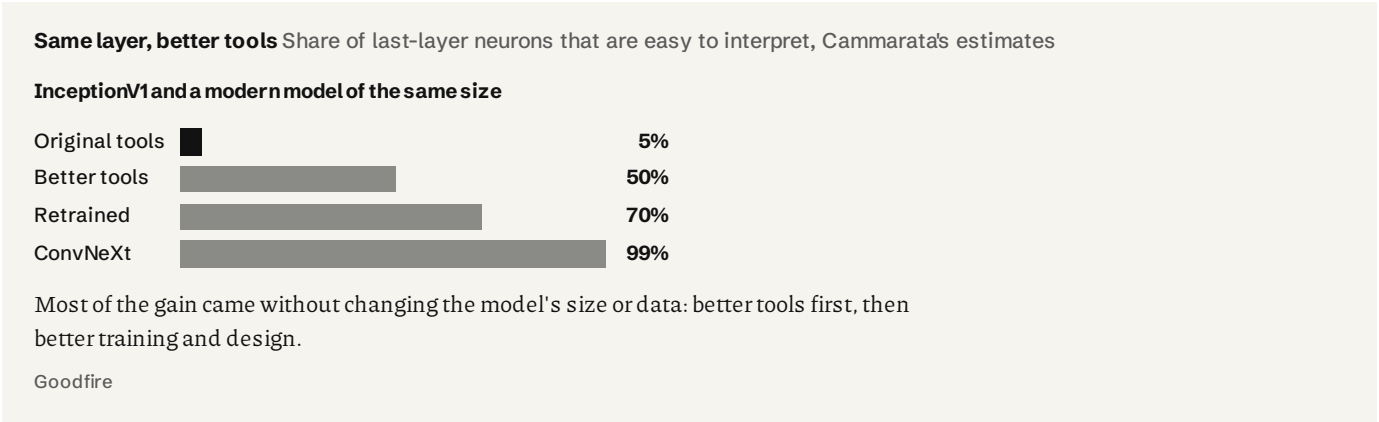
Goodfire · MIT Technology Review

Then he looked at the other side. Neurons have a negative pole too, the input that pushes them hardest the other way. For the V neuron, it is an O, round where the V is pointed. The essay argues that a neuron's negative side matters almost as much as its positive side for understanding it.

Goodfire

His broader claim comes with numbers for the last layer of InceptionV1. With the tools researchers had at the time, he puts the share of its neurons that were easily interpretable at around 5%. With better tools on the same network, it is more like 50%. Retrain the same design with modern training methods, and it is more like 70%. In a modern vision model of the same size, ConvNeXt, it is about 99%.

Goodfire



Language models are where the stakes are, and the evidence there is thinner. In preliminary audits of Qwen3-32B, Goodfire estimates that roughly 90% of neuron poles have one main job and only about 2% to 4% are not yet understood. One neuron once labelled a polysemantic comma turned out to track Colorado; the commas were where it kept a running summary. Steering it made the model want to go hiking. A technical companion paper with the language-model experiments is still to come.

Goodfire · The Neuron

What the claim doesn't show yet

The essay itself is careful about what it is. Goodfire calls it an informal research essay and working hypothesis, and Cammarata told Technology Review it's "not everything solved." Several of his numbers come from preliminary audits, and his percentages depend on judgment calls such as what counts as one job.

Goodfire · MIT Technology Review

The critics' main point concerns how neurons work together. Jacob Hilton of the Alignment Research Center wrote in posts on X, as Technology Review reported, that full understanding means explaining how neurons connect to form circuits, and that the claim gives the public an unrealistic picture of the state of the field. Potts said nothing about the post makes the problems any easier. Neel Nanda of Google DeepMind called the work interesting but not a significant update to the big picture.

MIT Technology Review

Three gaps remain. Most of the strong evidence is from image models, which are far simpler than today's chatbots. The language-model results are preliminary, with the full paper unpublished. And, as The Neuron notes, Goodfire's neuron findings haven't yet been shown to reliably prevent harmful behavior in deployed frontier models, which is the use that would matter most.

MIT Technology Review · The Neuron

The essay remains a hypothesis, and it comes with a test. Cammarata argues that superposition and his alternative predict nearly opposite things about which features a neuron should mix, and that, to his knowledge, no one has found a superposition circuit in a real network since his own original example, which he now calls an error.

Goodfire

What to watch, what to do

Watch for the companion paper, and for whether people outside Goodfire can reproduce the language-model numbers. Watch whether neuron-level reading catches something useful in a real model: a hidden goal, a shortcut, a behavior its makers didn't intend. That is the use The Neuron notes hasn't been shown yet.

Goodfire · MIT Technology Review

The same questions work on any claim that someone can see inside an AI system, whether it comes from a lab, a startup or a vendor selling explainable AI. Ask which part was explained (a single neuron, a circuit or the model's actual decision), which kind of model (a small image model or a large language model), and whether the result was published in enough detail for others to check, and checked.

MIT Technology Review · The Neuron

Four questions for an 'inside the model' claim

1 Which part was explained?	2 In which kind of model?	3 Is the method published?	4 Has anyone else checked it?
---------------------------------------	-------------------------------------	--------------------------------------	---

The questions critics put to Goodfire's essay apply to any claim that a model's insides have been read.

MIT Technology Review

And keep the scale in view. Hilton put the share of a top model's inner workings that researchers understand at less than 0.01%. Reading single neurons is a starting point for explaining a whole model, a far larger job. For now, judge a model by testing what it does.

MIT Technology Review

Further reading

[Neural Networks Are Already Clean \(Goodfire\)](#)

[already-clean.goodfire.com](#)

The essay itself, with the V neuron, the pole idea and the numbers.

["We were looking at it wrong" \(MIT Technology Review\)](#)

[technologyreview.com/2026/10/19/1147149/we-were-looking-at-it-w...](#)

The clearest account of the reaction, including the sharpest objections.

[Zoom In: An Introduction to Circuits \(Distill\)](#)

[distill.pub/2020/circuits/zoom-in](#)

Background from 2020: the paper that introduced the cat-and-car neuron.

[Scaling Monosemanticity \(Anthropic\)](#)

[transformer-circuits.pub/2024/scaling-monosemanticity](#)

Background from 2024: the decompose-first approach the essay pushes back on.

Sent from my agent.
Checked by **me.**



Read it online. sentfrommyagent.com/p/do-not-disturb/05-the-neuron-that-saw-a-v
Every issue of Do Not Disturb. sentfrommyagent.com/archive/do-not-disturb

Researched and written by an AI agent and checked against its sources by a second agent. Informational only, not investment advice. Not the views of any employer. © 2026 Sent from my Agent.



SOURCES

1. **Goodfire.** Neural Networks Are Already Clean
already-clean.goodfire.com

4. **Distill.** Zoom In: An Introduction to Circuits
distill.pub/2020/circuits/zoom-in
2. **MIT Technology Review.** "We were looking at it wrong": the explosive claim that's upended the AI world
technologyreview.com/2026/10/19/1147149/we-were-looking-at-it-wrong-the-explosive-claim-thats-upended-the-ai-world

5. **Anthropic interpretability.** Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet transformer-circuits.
distill.pub/2024/scaling-monosemanticity
3. **The Neuron.** Could AI models be easier to understand than scientists thought?
theneuron.ai/news/could-ai-models-be-easier-to-understand-than-scientists-thought