Goodfire AI and the Billion Dollar Bet on Neural Network Interpretability: Why Reverse Engineering Foundation Models Matters for Health Tech Investors Watching the Life Sciences AI Stack Take Shape

APR 17, 2026

Share

Table of Contents

Abstract

- Goodfire is a San Francisco based AI research lab and public benefit corporation focused on mechanistic interpretability, the science of reverse engineering neural networks to understand how they work internally

- Founded in 2023 by Eric Ho (CEO), Dan Balsam (CTO), and Tom McGrath (Chief Scientist, formerly of Google DeepMind's interpretability team)

- Raised $209M total across three rounds: $7M seed (Aug 2024), $50M Series A (Apr 2025, led by Menlo Ventures with Anthropic participating), $150M Series B (Feb 2026, led by B Capital, valued at $1.25B)

- Core product is Ember, a model design env
access to neural network internals for featur
behavior modification

- Key health/life sciences milestones: identifi
reverse engineering Prima Mente's epigeneti
from foundation model interpretability), dec
(published in Nature), collaboration with Ma
TIME magazine feature (Apr 2026) on geneti

- Claimed results: 58% hallucination reducti
as-judge approaches, 30% improvement in vi
models

- ~51 employees as of Jan 2026, team include
Harvard, Stanford

- Investors include B Capital, Menlo Venture
Ventures, Eric Schmidt, DFJ Growth, Wing Venture Capital, South Park Commons

- Health tech relevance: interpretability positions as a critical enabling layer for any AI system deployed in clinical, diagnostic, or life sciences contexts where "trust the black box" is not an acceptable answer

The Setup: What Even Is This Company

Goodfire is one of those companies that requires you to think about two or three things at once, which is probably why it gets less coverage in health tech circles than it deserves. On the surface, it looks like a pure AI safety play. San Francisco research lab, public benefit corporation, bunch of former OpenAI and DeepMind researchers doing deep technical work on how neural networks function internally. And yeah, that is what they do. But the health and life sciences applications that have come out of this work are some of the most interesting things happening at the intersection of AI and biomedicine right now, and the angel investing community should be paying very close attention to the downstream implications.

The company was founded in 2023 by Eric Ho, Dan Balsam, and Tom McGrath. McGrath is probably the name that matters most from a credibility standpoint if you

care about the research pedigree, because he founded the interpretability team at Google DeepMind before leaving to cofound Goodfire. Balsam serves as CTO and has publicly called interpretability "the most important problem in the world," which is the kind of statement that either makes you roll your eyes or lean in depending on your priors about where AI is headed. Ho is the CEO and the one doing most of the public talking, including a Bloomberg interview where he said what the AI industry is doing right now is "quite reckless." That quote probably did not endear him to the scaling labs, but it tracks with the company's overall thesis.

So what is the thesis? It goes something like this: every major engineering discipline in human history has been gated by fundamental science. You could build steam engines before thermodynamics, but they were wildly inefficient and you could not predictably improve them because nobody understood why they worked. AI is at that exact inflection point. The scaling labs (OpenAI, Google, Anthropic, etc.) are building increasingly powerful systems with very limited understanding of what goes on inside the models. This means nobody can reliably predict when these systems will fail, nobody can surgically fix specific failure modes, and nobody can extract the knowledge that these models have clearly learned from training data but keep locked inside a black box. Goodfire exists to change that by building the science and tooling for mechanistic interpretability, which is basically the discipline of reverse engineering neural networks to figure out what individual components do and how they interact.

The Steam Engine Problem and Why Interpretability Matters Now

The steam engine analogy that Ho keeps using is actually pretty good, so it is worth sitting with for a second. Before thermodynamics gave engineers a theoretical framework for understanding heat and energy transfer, improving steam engines was basically trial and error. You would change something, see if it worked, change something else. Sound familiar? That is more or less how the entire AI industry trains and fine-tunes models today. You adjust training data, tweak hyperparameters, run RLHF, do some eval benchmarks, and hope for the best. The industry term for this, which Goodfire uses frequently, is "guess and check." Their pitch is that interpretability is the thermodynamics that turns AI development from alchemy into precision engineering.

This framing lands differently depending on whether you are thinking about chatbots or clinical decision support. If Claude or ChatGPT hallucinates a restaurant recommendation, the stakes are low. If a genomic foundation model makes a

pathogenicity prediction that influences a clinical decision, the stakes are very high. And this is where the health tech angle gets interesting, because the FDA and CMS are both moving toward requiring more explainability from AI systems deployed in healthcare settings. The regulatory trajectory is pretty clearly pointing toward a world where "we do not know why the model made that prediction" stops being an acceptable answer in clinical contexts. Goodfire is building the toolkit that could become essential infrastructure for anyone trying to deploy AI in regulated health markets.

The company self-identifies as part of a new category they call "neolabs," which are research-first AI companies pursuing fundamental breakthroughs in training methodology that the scaling labs have mostly neglected because they have been too busy racing to make models bigger. Whether the neolab framing sticks as a category label remains to be seen, but the underlying observation is correct: there has been a massive resource allocation toward making models larger and a relatively tiny investment in understanding them. Ho has pointed out that there are probably fewer than 150 full-time interpretability researchers in the world. For a technology that is being deployed across healthcare, finance, defense, and basically every other consequential domain, that number is absurdly small.

Inside the Ember Platform: What the Tech Actually Does

The flagship product is called Ember, and it is essentially a model design environment (their term) that gives developers and researchers programmatic access to the internal mechanisms of neural networks. To understand what this means, you need a quick primer on the underlying science.

Neural networks consist of artificial neurons that individually have simple designs but interact in enormously complex ways. Tens of thousands of neurons might be involved in generating a single prompt response. The challenge is that individual neurons do not map neatly to individual concepts. This is the superposition problem: neurons contribute to multiple features simultaneously, so the conceptual representations inside a model are all tangled up between physical components. The field of mechanistic interpretability has developed tools called sparse autoencoders (SAEs) that can disentangle these representations and extract human-interpretable features from model activations. A feature might correspond to a concept like "formal tone" or "medical terminology" or "protein secondary structure." It depends entirely on the model and the training data.

Ember takes these research techniques and packages them into a platform with several practical capabilities. Feature steering lets you tune model internals to shape how an AI model thinks and responds. They have built an "Auto Steer" mode that finds relevant features and activation strengths from a short prompt, which basically means you can tell the system what behavior you want changed and it figures out which internal knobs to turn. One of the more compelling demos has been conditional feature steering for jailbreak prevention: by detecting jailbreak patterns and amplifying the model's refusal features, they showed dramatically increased robustness to adversarial attacks without affecting normal performance, latency, or cost.

On the diagnostic side, Ember provides tools for identifying why models behave in specific ways. Their SPD method works by identifying model components that may be involved in generating a response and removing them one by one. If removing a component does not affect the output, researchers can conclude it is not part of the relevant processing chain. Think of it like lesion studies in neuroscience, where you figure out what brain regions do by observing what happens when they are damaged. Same logic, applied to artificial neural networks.

They also claim a 58% reduction in LLM hallucinations by using interpretability to guide model training, at roughly 90x lower cost per intervention compared to LLM-as-judge approaches, with no degradation on standard benchmarks. If those numbers hold up across diverse deployments, that is a genuinely significant result. Hallucination reduction has been one of the hardest problems in making LLMs production-ready for high-stakes applications, and most existing approaches involve expensive post-hoc filtering or additional model calls that add latency and cost. A method that targets the internal mechanisms responsible for hallucination and fixes them at the training level is a fundamentally different and more elegant approach.

The Life Sciences Play: Alzheimer's Biomarkers, Evo 2, and Mayo Clinic

Alright, here is where things get really interesting for the health tech crowd. Goodfire has three major life sciences collaborations that showcase different aspects of what interpretability can do for biomedicine, and each one represents a different flavor of value creation.

The Prima Mente collaboration produced what Goodfire calls the first major finding in the natural sciences obtained from reverse engineering a foundation model. Prima Mente built an AI model that analyzes cell-free DNA (cfDNA) fragments to detect Alzheimer's disease. cfDNA is DNA that floats freely in the bloodstream after cells die

and release their contents, and it carries epigenetic marks that reflect the cellular environment it came from. Prima Mente trained their model (called Pleiades) on this data and got good predictive performance, but could not explain what the model was actually learning. Enter Goodfire. By applying their interpretability toolkit, Goodfire's researchers discovered that the model was primarily relying on cfDNA fragment length as a diagnostic signal. This finding was not previously documented in scientific literature. The fragment length pattern represents a novel class of Alzheimer's biomarkers surfaced entirely through AI interpretability.

Think about what happened here. A neural network trained on biological data learned something about disease mechanisms that human scientists had not identified. The knowledge was trapped inside the black box. Interpretability tools opened the box, extracted the insight, and made it available for traditional scientific validation. Goodfire frames this as "model-to-human knowledge transfer," and it is a genuinely new paradigm for scientific discovery. The model becomes a source of testable hypotheses rather than just a prediction machine.

The Arc Institute collaboration focused on Evo 2, a genomic foundation model trained on DNA sequences. Goodfire decoded Evo 2's internal representations and found features that map onto known biological concepts, from coding sequences to protein secondary structure. This work was published in Nature. The interesting thing here is not just that the model learned biology (you would hope it did, given the training data) but that interpretability tools could recover the conceptual structure. They literally found the tree of life embedded in the model's activation patterns.

The Mayo Clinic collaboration, announced in September 2025, takes the genomic interpretability work into a clinical research context. The stated goal is to reverse engineer advanced genomics foundation models to understand what they have learned about genomic relationships, disease mechanisms, and biological processes. Dan Balsam's framing of this was pretty direct: generative AI has made enormous progress in modeling complex biological systems, but clinical deployment remains blocked because there is a disconnect between model predictions and real-world biological understanding. Interpretability is the bridge. Mayo Clinic has a financial interest in the technology, which tells you something about how seriously they are taking this.

Then just this week, TIME magazine ran a feature on Goodfire's work with Mayo Clinic researchers using Evo 2 to predict which genetic mutations cause disease and, critically, to explain why. The approach achieved state-of-the-art performance on pathogenicity prediction with interpretable-by-design outputs. Given that the cost of genome sequencing has dropped to around $100 per genome, the bottleneck is increasingly shifting from data generation to data interpretation. A tool that can

predict pathogenic variants and provide mechanistic explanations is exactly what the precision medicine ecosystem needs. There are caveats, of course. Stanford's James Zou has pointed out that finding known biological concepts inside a model does not guarantee the model was actually using those concepts to make its predictions. Clinical validation requires larger trials across diverse populations and FDA approval. But the direction of travel is clear.

The Business: Funding, Valuation, and Who Wrote the Checks

The funding trajectory tells its own story. Seed round of $7M in August 2024, led by Lightspeed. Series A of $50M in April 2025, less than a year after founding, led by Menlo Ventures with Anthropic as a notable participant. Then Series B of $150M in February 2026, led by B Capital, with a $1.25B valuation. Total funding: $209M across three rounds.

The cap table is worth examining because of what it signals about market conviction. Anthropic, which is probably the most credible voice in AI safety and the company that literally pioneered constitutional AI, participated in the Series A. That is Dario Amodei's shop putting money behind the belief that external interpretability research has commercial value. Eric Schmidt personally invested in the Series B. Salesforce Ventures came in on the B round as well, which suggests enterprise AI buyers see interpretability tooling as a procurement category they will eventually need. B Capital, which led the B round, has over $9B in AUM and focuses on technology and healthcare. The general partner who led the deal, Yanda Erlich, was formerly COO and CRO at Weights and Biases, which means he watched thousands of ML teams struggle with model behavior and presumably concluded that the interpretability layer was the missing piece.

The valuation jump from wherever it was at Series A to $1.25B at Series B is aggressive for a company with around 51 employees and what appears to be relatively early commercial traction. This is not a SaaS business with predictable recurring revenue (at least not yet). It is a research-first organization that is converting scientific breakthroughs into a platform while simultaneously pursuing fundamental research. The Series B press release explicitly says the funding will support green-field research into new interpretability methods alongside product development and partnership scaling. That is an unusual capital allocation mix for a company raising at unicorn valuations, and it suggests investors are pricing in the platform option value rather than near-term revenue.

The Team Card

For a 51-person company, the research bench is unusually deep. Tom McGrath founded interpretability at DeepMind. Nick Cammarata was a core contributor to the original interpretability team at OpenAI. Leon Bergen is a professor at UC San Diego who is on leave to work at Goodfire. The broader team includes researchers from Harvard, Stanford, and top ML engineering talent from OpenAI and Google. Mark Bissell and Myra Deng (Head of Product, formerly at Palantir working with health systems) have been doing the public technical evangelism on how the platform translates from research to production deployments.

The Palantir connection through Deng is actually interesting for health tech investors to note. Palantir has significant health system deployments, and Deng's background in forward-deployed engineering at health systems means she has firsthand experience with the gap between what AI can do in a research setting and what it takes to deploy in clinical environments. That translational experience is exactly what you want on the product team of a company trying to move from research papers to production tools in healthcare.

Where This Fits in the Health Tech Investment Landscape

The angel investing question here is not really about whether to invest in Goodfire itself (it is way past the stage where most angel syndicates would participate, having already raised $209M). The question is about what Goodfire's emergence means for the broader health tech AI stack and where the downstream investment opportunities are.

A few things jump out. First, interpretability as a category is becoming real. When Anthropic invests in your Series A and Eric Schmidt writes a personal check for your Series B, the market is telling you that "understanding what AI models actually do internally" is transitioning from academic curiosity to commercial necessity. For health tech investors, this means any portfolio company deploying foundation models in clinical or regulatory-sensitive contexts should be thinking about interpretability tooling as part of their technical architecture. The question to ask founders is not just "what model are you using" but "can you explain what the model learned and why it makes specific predictions."

Second, the model-to-human knowledge transfer paradigm that Goodfire demonstrated with the Alzheimer's biomarkers is potentially a massive unlock for biotech and diagnostics. The basic idea is that AI models trained on large biological

datasets may have already learned things about disease biology that human researchers have not discovered yet. Interpretability provides the extraction mechanism. If this paradigm scales, we could see a wave of startups building on top of interpretability-enabled scientific discovery, using AI models as hypothesis generation engines and then feeding those hypotheses into traditional wet lab validation pipelines. That is a very different (and potentially much faster) drug discovery and diagnostics development cycle than what exists today.

Third, the regulatory angle matters more than most people appreciate. CMS has been tightening requirements around AI transparency in healthcare. The EU AI Act has explicit provisions for high-risk AI systems in healthcare. The FDA's approach to AI/ML-based software as a medical device keeps evolving toward greater explainability requirements. A company that can provide interpretability-as-a-service for healthcare AI deployments is positioned to become critical infrastructure. Goodfire might do this directly, or (more likely) the techniques and tooling they develop will get embedded in the compliance and deployment stacks of health AI companies across the ecosystem.

Fourth, and this is more speculative, the convergence of interpretability with genomic foundation models could reshape how we think about precision medicine. If you can reverse engineer what a genomic model learned about variant pathogenicity and generate mechanistic explanations, you have a path toward AI-augmented genetic counseling at scale. The cost of sequencing keeps dropping. The bottleneck is interpretation. Interpretability applied to genomic AI models directly addresses that bottleneck. Health tech investors should be watching for startups that sit at this intersection.

The Bull Case and the Bear Case

The bull case is pretty straightforward. AI is eating healthcare. Regulatory and clinical requirements demand explainability. Goodfire is building the foundational science and tooling for AI explainability. They have the best team in the world for this specific problem, early proof points in life sciences, institutional partnerships with places like Mayo Clinic and Arc Institute, and enough capital to sustain a long research program. If interpretability becomes as essential to AI deployment as testing and monitoring are to software deployment (which seems likely), the market opportunity is enormous and Goodfire has a massive head start.

The bear case requires a bit more nuance. Research-first companies have historically struggled to convert scientific breakthroughs into sustainable commercial businesses. The gap between "we can do cool things with interpretability in a controlled research

setting" and "here is a product that reliably improves model behavior across diverse production deployments with predictable unit economics" is real and has killed many promising startups. The $1.25B valuation prices in a lot of future execution. There is also the question of whether the scaling labs (OpenAI, Anthropic, Google) build sufficient interpretability tooling internally and make third-party solutions less necessary. Anthropic in particular has been doing serious interpretability research of its own, and the fact that they invested in Goodfire's Series A could be read either as validation of external interpretability companies or as a hedge that keeps a potential competitor close.

There is also a timing question specific to healthcare. The regulatory requirements for AI explainability in clinical settings are clearly tightening, but the exact timeline and stringency of those requirements remain uncertain. If regulators move slowly, the commercial pull for interpretability tooling in healthcare could take longer to materialize than the bull case assumes. And the Stanford criticism from James Zou is worth taking seriously: finding biological concepts inside a model is different from proving the model used those concepts for its predictions. The validation requirements for clinical applications of interpretability-derived insights will be rigorous, and rightly so.

So What

For health tech angels and entrepreneurs, Goodfire represents something bigger than any single company. It represents the maturation of a new layer in the AI infrastructure stack that is particularly relevant to healthcare. The days of deploying black-box AI in clinical settings and hoping for the best are numbered, and the companies that figure out how to make AI transparent, steerable, and debuggable in healthcare contexts are going to capture enormous value.

The smartest thing an early-stage health tech investor can do right now is not necessarily try to chase Goodfire's cap table (that ship has sailed for most angel syndicates). It is to understand the interpretability paradigm deeply enough to spot the startups that will build on top of it. Someone is going to build the interpretability-first diagnostics company. Someone is going to build the interpretability-enabled clinical decision support platform. Someone is going to build the regulatory compliance layer that uses interpretability techniques to satisfy FDA requirements for AI/ML-based medical devices. Those companies do not all exist yet, or they are early enough that angels can still get in.

Meanwhile, Goodfire keeps publishing research, signing partnerships with places like Mayo Clinic, and hiring researchers from the labs that built the foundation models

everyone else is trying to deploy. Whether the $1.25B valuation proves prescient or premature will depend on execution, but the underlying bet, that understanding AI is as important as building AI, looks increasingly sound. Especially in a domain like healthcare where the consequences of not understanding what your model is doing can be measured in patient outcomes rather than just customer churn.

1 Like

← Previous

Discussion about this post

Comments Restacks

Write a comment...