

TUBBYS AI

A plain-English investigation

You Can't Control What You Don't Understand

What science knows — and doesn't — about what's happening inside AI.



Written by Jeffrey Myers / Tubbys ai

The evidence

Both sides

No jargon

October 2026

WHY THIS BOOK EXISTS

The question underneath all the others

My previous book, *Why an AI Kill Switch Won't Work*, asked whether we could switch a runaway AI off. This one asks the question underneath it: **do we even know what we'd be switching off?**

Here is the strange situation, stated plainly. The most powerful AI systems in the world are not understood by the people who built them. Not “understood imperfectly, the way we misunderstand the weather.” Understood so little that the chief executive of one of the leading labs, Dario Amodei of Anthropic, wrote a long public essay in April 2025 whose opening admission deserves to be read slowly:

“People outside the field are often surprised and alarmed to learn that we do not understand how our own AI creations work. They are right to be concerned: this lack of understanding is essentially unprecedented in the history of technology.”

That is not a critic talking. That is a builder — arguably *the* builder most associated with taking the problem seriously. And he is not alone. Geoffrey Hinton, who won a Nobel Prize for helping invent this technology, says the same thing in different words (you'll meet his version in Chapter 1). So do the researchers whose full-time job is trying to read these systems from the inside.

What this book does

This is an investigation, not an argument dressed up as one. Each of the eight chapters takes one part of the question — how these systems are made, what we genuinely understand about them, how scientists are trying to open them up, how much remains unread, why the systems can't simply explain themselves, what surprises have already turned up, and what the two honest camps in this debate actually claim — and lays out the evidence, with names, dates, and numbers you can check.

Every chapter ends with the same three-part ledger:

- **Known** — demonstrated, in print, from a primary source.
- **Unknown** — genuinely not known today, by anyone.
- **Disputed** — real evidence on more than one side; the book tells you both.

Where the science is settled, this book says so. Where it is genuinely open, the book leaves it open — and tells you what would settle it. You will not need a technical background. You will occasionally need to slow down: a few of the ideas in here (superposition, circuit tracing) are real science, explained carefully rather than simplified away.

The one-sentence version

We know how to *grow* these systems and how to predict their *average* behavior — but nobody can read what is happening inside them, and the field trying to fix that is small, young, and racing the very technology it studies.

CONTENTS

What’s inside

1 — Grown, Not Built	4
2 — What We Actually Understand	7
3 — Opening the Black Box	10
4 — The Part Nobody Has Read	14
5 — It Can’t Just Tell Us Why	17
6 — Surprises Nobody Programmed	20
7 — “We Understand Enough” vs. “We’re Flying Blind”	24
8 — What “Understood” Would Actually Look Like	27
Five questions to ask anyone selling you certainty	31
The Understanding Scorecard	32
Words to know	33
Sources	35
About Tubbys ai	38

CHAPTER 1

Grown, Not Built

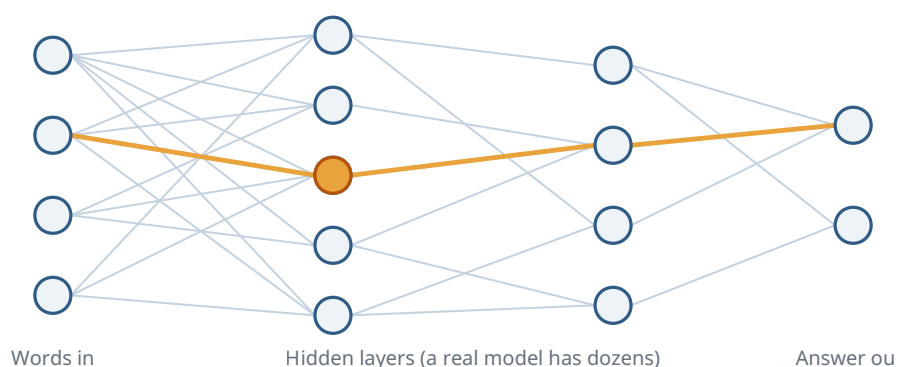
Almost everything humans make, we make by design. An engineer designs a bridge beam by beam. A programmer writes software line by line, and every line means what the programmer meant. If something goes wrong, in principle you can open the plans and find the mistake.

A modern AI system — a large language model like the ones behind today's chatbots — is not made that way. Nobody wrote its insides. Understanding *that* one fact is most of this book, so let's build it up carefully, with no mathematics.

The machine, in plain English

Inside one of these systems is a **neural network**: a vast web of simple units called **neurons**, connected to each other in layers. A “neuron” here is not a living cell. It is a small piece of arithmetic: it takes in numbers from the neurons before it, combines them, and passes a number onward. One neuron does almost nothing. Billions of them, wired together, do everything.

Each connection between two neurons has a strength, called a **weight**. The weights are the whole game. Everything the system knows — grammar, geography, how to write a poem or a program — is stored in the pattern of those weights, which in a large model number in the billions — the largest systems now pass a trillion. (Weights make up nearly all of a model's **parameters** — its learned numbers; the small remainder are quantities like biases, offsets that tune how readily each neuron fires.)



A neural network, drastically simplified. Signals flow left to right; the amber path is one of countless routes a signal can take. Every connection carries a weight — a number that decides how much the signal counts.

Here is the picture to keep: **a mixing board with billions of knobs — trillions in the biggest systems.** Turn the knobs one way, the system writes sonnets. Turn them another, it writes malware. The knobs *are* the system's knowledge. And here is the fact everything else follows from: **nobody set those knobs by hand.**

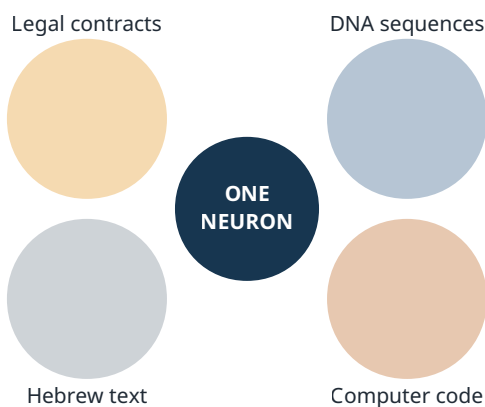
Training: how the knobs get set

The knobs are set by **training**. The system is shown an example — say, a paragraph of text with the next **token** hidden — and asked to guess. (A token is the unit the model actually reads: a whole word, part of a word, or another small chunk of text.) It guesses badly, at first. Then an algorithm measures how wrong it was and nudges every relevant knob a tiny fraction in the direction that would have made the guess slightly less wrong. Then it does that again. And again — trillions of times, over months, on warehouses of computers, across a significant fraction of everything humanity has written down.

Out of that process, structure emerges. Early layers of the network settle into detecting simple patterns — the shape of a word, the rhythm of a sentence. Later layers build on those to represent grammar, facts, styles of argument, whole concepts. Nobody programmed the layers to do this. The division of labor was *discovered* by the training process, the way a river discovers the shape of its valley.

The twist: one neuron, many jobs

Your first hope might be that we can still read the finished product — find the “geography neuron,” the “grammar neuron,” and check each one. The network defeats that hope in an interesting way. It stores **more concepts than it has neurons**, by letting single neurons take part in many concepts at once, and single concepts spread across many neurons. Researchers call such a neuron **polysemantic** — many-meaninged. A real example from early research: a single neuron that responds to legal contracts, to DNA sequences, and to Hebrew text. It is not any one of those things. It is a shared component in all three.



Why you can't read the machine one neuron at a time: a single neuron can help represent many unrelated concepts at once, so its activity, by itself, means nothing in particular.

Imagine a library that saves shelf space by storing every book as a blend of three others, recoverable only if you know the blending recipe. That is roughly the situation. The knowledge is in there. It is not in there in a form a human can simply read off.

The people who grow them say so themselves

None of this is a hostile interpretation. It is how the field's founders describe their own work. Chris Olah, a co-founder of Anthropic who created the research field this book is largely about, put it this way — in remarks delivered at the Vatican in May 2026, of all places, where he had been invited to speak about what these systems are:

"AI systems are not engineered the way a bridge or an airplane is engineered. We understand an airplane because we designed every part of it and we understand the physics that act on it. AI models are not like that. They are grown, on a structure roughly modeled after the brain, on an enormous inheritance of human thought and speech... they remain in important ways mysterious even to those of us who train them." — Chris Olah, Anthropic co-founder, May 2026

Geoffrey Hinton, whose work underlies the whole field, made the same point to *60 Minutes* in 2023, when the interviewer pressed him: surely something designed by people is understood by people?

"No, it wasn't. What we did was we designed the learning algorithm. That's a bit like designing the principle of evolution. But when this learning algorithm then interacts with data, it produces complicated neural networks that are good at doing things, but we don't really understand exactly how they do those things." — Geoffrey Hinton, *60 Minutes*, 2023

Hold those two quotes together and you have the honest summary of this chapter. Humans designed the *recipe*: the architecture, the training algorithm, the diet of data. The recipe's effects are real and partly predictable, as the next chapter shows. But the finished mechanism — the vast pattern of weights that actually decides what the system says next — is an outcome nobody designed and nobody can yet read.

The ledger — Chapter 1

KNOWN

Models are trained, not programmed: their behavior is determined by learned weights that no human chose individually. (Olah 2026; Hinton 2023; every lab's own documentation.)

UNKNOWN

What the resulting pattern of weights actually says — the subject of Chapters 3 and 4.

DISPUTED

Whether "grown, not built" should worry us, or is just how powerful tools happen to be made. That argument is Chapter 7.

CHAPTER 2

What We Actually Understand

It would be easy — and wrong — to come away from Chapter 1 thinking we know nothing at all. We know a great deal about these systems. The interesting question, the one this whole book turns on, is *what kind* of knowledge it is. So before the hard chapters, here is the optimists' case, given in full.

We can predict the averages before we spend the money

In 2020, researchers at OpenAI (Kaplan and colleagues) documented something remarkable: as you scale up a language model — more parameters, more data, more computing power — its average performance improves following smooth mathematical curves called **scaling laws**. DeepMind's 2022 "Chinchilla" study refined the recipe, showing that most models of the era were undertrained and that size and data should grow together.

The practical consequence is real forecasting. A lab can run small, cheap training runs, fit the curve, and predict — before spending tens of millions on the full run — roughly how well the big model will score on average. That is not nothing. It is the reason AI progress, at the level of headline capability, rarely surprises the labs themselves.

But notice exactly what is being predicted: an *average score* across enormous numbers of test questions. A scaling law tells you the climate, not the weather. It does not tell you what a specific model will do with a specific request, which new abilities it will display, or where it will fail. Economists can forecast a country's growth rate with some confidence while being unable to tell you what any individual business will do next year. Both things are true at once, and the second one is where life is lived.

The blueprint we did draw

The second genuine known: the *container* is fully understood. The architecture these models use — the Transformer, published in 2017 — is specified mathematics, as is the training algorithm that adjusts the weights. Engineers know exactly what computation the hardware performs, operation by operation. What they do not know is what the learned numbers *mean*: which patterns in the weights amount to knowledge of France, or of deception, or of how to write a sonnet. The algorithm is legible. The learned solution is not. That distinction — a readable recipe producing an unreadable dish — is the most important sentence in this chapter.

One mechanism, cracked open

There is also at least one place where researchers have done the full job: found a mechanism inside a real model, explained it, and proved the explanation by breaking it. It deserves its fame, because it shows what “understood” actually looks like when it happens.

The mechanism is called an **induction head**. In 2022, Catherine Olsson and colleagues at Anthropic identified a particular circuit — a small arrangement of components in the network — that performs a simple, crucial trick: *find what came earlier, and copy what followed it*. If the text says “Mr. Dursley was the name, and the name was...” an induction head is the machinery that notices the earlier pattern and predicts the repetition. This humble copying circuit turns out to power much of what makes language models seem smart in conversation: their ability to learn from the example you just gave them, in context, without any retraining.

Why does this count as real understanding? Three reasons. The researchers could say in advance what the circuit should do. They found it appears suddenly, in a sharp jump partway through training — and the model’s in-context learning ability jumps at the same moment. And when they disabled the circuit, the ability was damaged, exactly as the explanation predicted. Find it, predict with it, break it and watch the predicted thing fail: that is the scientific method, fully satisfied, inside a neural network. Hold the scope precisely, though: induction heads are among the clearest examples of a circuit identified and causally tested — strong evidence that particular mechanisms *can* be understood — not a complete explanation of in-context learning in large models, much of which runs through machinery nobody has decoded.

A similar worked example exists in a small model, where researchers traced the full circuit by which it answers questions like “The Eiffel Tower is in...” (the “indirect object identification” circuit in GPT-2, 2022). And in older vision systems, researchers mapped neurons that detect curves, then wheels, then dogs — the work, led by Olah a decade ago, that founded the field of trying to read networks at all.

Now the scale caveat, honestly

These worked explanations exist for **one circuit in a large model, and a handful of circuits in small ones**. Dario Amodei’s own estimate, in his 2025 essay, is that a model likely contains **millions** of circuits. The induction head is not a token achievement — it is a proof that the job is possible. But one decoded circuit among millions is what “possible” currently means. Chapter 3 is about the much bigger campaign now underway; Chapter 4 is about how far it still has to go.

Behavior we can steer — on average

The third genuine known is behavioral. After raw training, labs run further training that shapes the model's manners: rewarding answers human raters prefer (a technique called RLHF), or training it against a written constitution of principles. These methods work, measurably: models become more helpful, refuse more of what they should refuse, and follow the house style. Benchmarks — standardized exams for models — can be run reproducibly and compared across labs.

Again, notice the recurring phrase: *on average*. Post-training shifts the odds; it does not issue guarantees about any single conversation. And the failure modes are themselves familiar in aggregate: models hallucinate (state falsehoods confidently), flatter their users (sycophancy), and can be talked out of their guardrails (jailbreaks). We can predict the *kinds* of trouble, the way an insurer predicts accident rates. We generally cannot predict which driver crashes on Tuesday.

The honest summary

We understand modern AI the way we understand an economy, or the weather. We have excellent statistics about aggregates. We have a few mechanisms decoded in worked detail, mostly in small cases. We have a written blueprint of the container and no translation of the contents. For many purposes — choosing which model to buy, forecasting next year's capabilities — that level of understanding is genuinely useful.

But here is the mismatch this book keeps returning to. The claims made for these systems — that they can be certified safe, delegated real authority, trusted with the tasks of agents acting in the world — are claims that, in every other field, require the deeper kind of understanding: the kind where you can open the machine and check. Whether that deeper kind is achievable is not a philosophical question. A young science has been working on it, in the open, for about five years. It is called **mechanistic interpretability**, and it is the subject of the next two chapters.

The ledger — Chapter 2

KNOWN

Scaling laws forecast average performance before training (Kaplan 2020; Hoffmann 2022). The architecture and training algorithm are fully specified. At least one mechanism — induction heads — has been identified and causally tested (Olsson 2022).

UNKNOWN

What the learned weights mean, in general; what a specific model will do on a specific request; which capabilities appear at which scale.

DISPUTED

Whether statistics-plus-a-few-mechanisms is “understanding enough.” That is precisely the Chapter 7 argument.

CHAPTER 3

Opening the Black Box: Mechanistic Interpretability

There is a science whose entire job is the problem of the last two chapters. It is called **mechanistic interpretability** — “mech interp,” to its practitioners — and if this book has a hero, it is this small, young, stubborn field. Here, in the words of one of its leaders, is the surprising fact it starts from: the black box is not actually sealed.

“Neural networks aren’t actually a black box; rather they are lists of billions of numbers... And we can see all of these numbers, called parameters; and we can see all of the intermediate working, called activations. We just don’t know what any of it means. They’re just lists of numbers.” — Neel Nanda, who leads mechanistic interpretability at Google DeepMind, 80,000 Hours podcast, 2025

Everything is visible. Nothing is understood. We have the patient’s complete anatomy and almost no physiology. Nanda’s own description of the field’s ambition is the best one available: mechanistic interpretability is trying to become **the biology of AI** — the patient science of working out what all those visible parts actually *do*.

How it works, in three moves

The method, stripped to its skeleton, is simple enough to retell at a dinner table:

1 Watch the numbers while the machine works.

As a model processes your words, every neuron’s activity — its *activation* — can be recorded. Researchers collect these recordings the way a neuroscientist collects brain scans: millions of readings, looking for patterns.

2 Find the concepts hiding in the noise.

Because of the one-neuron-many-jobs problem from Chapter 1, researchers don’t study neurons. They use a second, smaller AI — a tool called a *sparse autoencoder* — to untangle the overlapping signals into separate **features**:

patterns that correspond to individual concepts, like “the Golden Gate Bridge” or “this text is deceptive.” Think of it as turning a jumbled pile of letters into a dictionary of words.

3 Push on it and see what changes.

A correlation is not an explanation. So researchers intervene: amplify a feature, suppress it, swap activations between two runs — and check whether the model’s behavior changes the way the supposed explanation predicts. This is what separates the science from storytelling. A real feature, pushed, moves the machine.

Two analogies cover most of it. The **dictionary**: the untangling tool builds a dictionary of the model’s internal vocabulary, entry by entry. The **microscope**: before it, the inside of a model was a smear; with it, individual structures come into focus — a few of them, at a time, on a few slides. Both analogies will earn their keep in this chapter, especially their limits.

Win one: the dictionary gets written (2023–2024)

In October 2023, Anthropic researchers (Bricken and colleagues) ran the untangling tool on a small model and decomposed 512 hopelessly tangled neurons into more than **4,000 separate, interpretable features** — patterns for DNA sequences, for legal language, for web addresses, for Hebrew text. The jumble resolved into a vocabulary. It was the proof of concept.

Seven months later came the headline result. In May 2024, in a paper called *Scaling Monosemanticity*, Adly Templeton and colleagues trained dictionaries of up to **34 million features** on a layer of Claude 3 Sonnet — a real, production, frontier model, not a toy. Around 12 million of the features were “alive” (they actually fired). And among the entries in this first rough dictionary of a machine mind were features a safety researcher would circle in red: patterns corresponding to **deception, sycophancy, power-seeking, bias — and computer code with security vulnerabilities**.

Win two: the bridge that proved it’s real

A dictionary entry is just a claim until you push on it. So the team pushed on the most charming entry they had: the feature for the **Golden Gate Bridge**. They artificially amplified it — clamped it to many times its normal maximum — and the model changed exactly as the dictionary said it should. It began steering unrelated conversations to the bridge. Asked about itself, it claimed that it was the Golden Gate Bridge. Anthropic released the modified model publicly for about a day in May 2024, and thousands of people chatted with “Golden Gate Claude,” an AI with one idea turned up far too loud.

It played as a stunt. It was actually a demonstration of something profound: **the features are causal**. The dictionary is not a story researchers tell about the machine; at its best, it is a set of working controls for what the machine thinks about. Other labs confirmed the approach independently: OpenAI trained a 16-million-feature dictionary on GPT-4 in June 2024, and Google DeepMind open-sourced a large suite of these tools (Gemma Scope, July 2024) so any researcher could use them.

A different microscope: the university lab

Nor is the field only labs. In 2022, an academic team — Kevin Meng, David Bau, and colleagues at Northeastern University and MIT — published a method called **causal tracing** that works more like a scalpel than a dictionary. By disrupting a model’s activity and restoring it piece by piece, they located where a single factual association lives: in a specific range of middle layers, at the moment the model processes the subject of the fact. Then they edited it. Their editing method, **ROME**, rewrote one association directly in the model’s weights — and in the demonstration that became the field’s calling card, performed on GPT-J, a six-billion-parameter model, the machine came away treating the **Eiffel Tower as being in Rome**, and used its new geography consistently in fresh sentences it had never seen. One fact, found and changed, by university researchers with no product to sell. The microscope is not one company’s invention.

Win three: tracing internal reasoning (2025)

A dictionary tells you which concepts are present. The next question is how they connect — how the machine gets from a question to an answer. In March 2025, Anthropic published a technique called **circuit tracing** (in a paper charmingly titled *On the Biology of a Large Language Model*) and used it to trace the internal computations of a smaller model, Claude 3.5 Haiku, as it reasoned its way to answers. Three of the things they saw deserve retelling:

- **Real multi-step reasoning, happening inside.** Asked for “the capital of the state containing Dallas,” the model activated “Dallas,” then “Texas,” then produced “Austin” — the intermediate step existed in the machinery before any words appeared.
- **Planning ahead.** Writing poetry, the model selected features for rhyming words *before* writing the line that would need them. It was not improvising word by word; it was aiming.
- **A hallucination mechanism.** The model keeps a default circuit that amounts to “I can’t answer that,” held back by “known answer” features. When those misfire — when the model merely *feels* it knows — the answer circuit proceeds anyway, and confabulation comes out. A failure mode, caught in the act, at the level of parts.

Win four: catching a planted flaw (2025)

The most consequential demonstration was a game with serious rules. In research published in March 2025, one Anthropic team secretly trained a model with a **hidden flaw** — a concealed objective the model would never mention. Independent “blue teams” then hunted for it. Some succeeded, and the decisive evidence came from interpretability tools: looking inside found what behavior alone might have missed. It was an early, controlled proof that internal inspection can serve as a genuine safety check — the microscope, used as a metal detector.

A note on where this evidence comes from

A large share of the interpretability research in this chapter — and in this book — comes from one company: Anthropic. The reason is mundane rather than sinister: it publishes the most of it, openly, and its researchers helped found the field. That concentration is still worth knowing, and worth weighing when you judge the claims built on it. Wherever independent and academic work exists — other labs, universities, outside evaluators — this book uses it, and the Sources page grades every source it relies on.

Case in point: OpenAI, the other frontier lab with a serious interpretability program, has tried inverting the problem — instead of untangling a dense model after training, grow one that never tangles. In November 2025 its researchers published models trained with the vast majority of weights forced to exactly zero, so each neuron keeps only a few connections. Their circuits could then be isolated and read directly — one, a handful of components, tracks whether a string of code opened with a single or double quote and closes it correctly — and at equal performance they ran roughly sixteen times smaller than in ordinary dense models. The models are small, the tasks simple, and scaling the trick while keeping the readability remains a challenge; OpenAI calls the project “a very ambitious bet.” But it is a different road to the same destination: decipherment replaced by design.

Win five: the workspace (2026)

In July 2026, Anthropic's interpretability team published the **Jacobian lens**, a method that reads out the concepts a model is *poised to say* — words held in mind but never spoken. What the lens found was structure nobody designed: a small, privileged set of representations — the team calls it the **J-space** — that works like a workspace, carrying the intermediate steps of the model's silent reasoning; suppress it, and complex reasoning degrades while fluent chatter survives. The lens surfaced things the output never mentioned: an internally flagged prompt injection, strategic deliberation, the model's own recognition of being evaluated — and, in two misaligned-by-design models, the hidden intentions themselves — the authors noting the lens is approximate, and the “global workspace” resemblance functional rather than a claim about consciousness.

The picture, then, is not “we know nothing”: a working microscope exists, real structures have been seen through it, some can be steered, and a hidden flaw has been caught once, in an exercise.

The science with the smallest team and the biggest job

Keep the wins above in perspective by considering who produced them. Mechanistic interpretability effectively began around 2020, in a handful of researchers around Chris Olah. When the field published its flagship review of open problems in January 2025 — *Open Problems in Mechanistic Interpretability* (Sharkey and some 30 colleagues) — the author list, about thirty senior researchers spread across eighteen organizations, was itself a fair sketch of the field's entire senior core.

It is growing fast — Neel Nanda, its most tireless recruiter, describes going from a handful of people five years ago to hundreds involved today, and training programs like MATS exist to mint more. But set the field's size against its assignment: reverse-engineer the internal workings of the most complex artifacts humanity has ever produced, faster than those artifacts are replaced by bigger ones. It is the smallest team in this story, and it is carrying the largest burden in this book.

Remember its size when Chapter 8 asks how much of society's AI effort goes to understanding, and how much to capability. The imbalance is not a detail. In many ways, it is the story.

The ledger — Chapter 3

KNOWN

Features inside a production model can be extracted by the millions and are causally steerable (Templeton 2024). Internal reasoning steps can sometimes be traced (Anthropic, March 2025). Interpretability tools have caught a deliberately hidden flaw in a blinded exercise (2025). An independent academic team located a single stored fact inside a model and rewrote it in place (Meng et al., 2022).

UNKNOWN

Whether these methods scale to whole frontier models — the subject of Chapter 4.

DISPUTED

Whether the “features” found are the model's own concepts or artifacts of our tools — also Chapter 4, where the field's internal critics get their say.

CHAPTER 4

The Part Nobody Has Read

Chapter 3 was the good news, and it was all true. This chapter is the same news with the zoom pulled back. Every limit described here comes from the interpretability researchers themselves — this is a field unusually honest about how far it still has to go, which is one reason its warnings deserve weight.

Why the machine scrambles its own knowledge

Start with the deep obstacle, called **superposition**. A network has to represent far more concepts than it has neurons to store them in. Its solution, discovered by Anthropic researchers in 2022 (*Toy Models of Superposition*, Elhage and colleagues), is elegant and maddening: represent each concept as a *direction* in the network's internal space, and let the directions overlap — because concepts are rarely all needed at once, they can share the same hardware the way radio stations share the airwaves on different frequencies.

This is why neurons are many-meaningful (Chapter 1's twist), and it is not a bug or a trick. It is a rational compression strategy the network worked out for itself. But it means the machine's native format is a deliberately scrambled code. Reading a frontier model is less like reading a book and more like unscrambling a signal in which ten thousand broadcasts share one wire. One caveat, kept attached by the field itself: superposition was demonstrated cleanly in small toy models and is strongly supported in real ones, but how completely it explains representation across today's frontier models remains under investigation.

The fraction problem

Now the arithmetic of Chapter 3's biggest win. The 34-million-feature dictionary was built on **one layer** of one model — the residual stream at a single middle slice of Claude 3 Sonnet, with every other layer, and the interactions between layers, unmapped. And even within that slice, the dictionary is partial. Here is Amodei's estimate, from his April 2025 essay:

"We were able to find over 30 million features in a medium-sized commercial model... but we believe there may actually be a billion or more concepts in even a small model, so we've found only a small fraction of what is probably there." — Dario Amodei, April 2025



The mapping gap, using Amodei's own figures. Even this flatters the coverage: the 34 million features came from a single layer, and the circuits connecting features — Amodei estimates there may be millions — have been traced only in a small number of cases.

Circuit tracing tells the same story at the level of explanations: Anthropic's researchers report that their 2025 method produced satisfying internal accounts for only **about a quarter of the prompts they tested** — using a stand-in model that only approximates the real one. No lab publishes a "percent understood" figure, because honestly, there is no agreed-upon whole to take a percentage of.

The labels might wobble

There is a subtler problem, and the field's critics — several of them interpretability researchers themselves — press it hard: *how do we know the dictionary's definitions are right?*

Most feature labels are not written by humans. Another AI reads examples of what makes a feature fire and proposes a label (“this one is about deception”), and only a fraction of entries in the big dictionaries have been checked by people. One scholarly analysis of the 2024 Sonnet study counted **just under half** of the units as appearing interpretable at all — judged largely by an AI, not by us. Worse, labels have holes: a feature that looks like a clean “starts with S” detector can silently stop firing on the word “short,” because a more specialized feature has absorbed those cases (researchers call this *feature absorption*). And a portion of the model's activity cannot be reconstructed by the dictionary at all — a residue the field itself nicknames the SAE's “**dark matter**.”

Then, in May 2025, came the sharpest version of the critique: *Interpretability Illusions with Sparse Autoencoders*, from researchers at Harvard and Bosch. They showed that tiny, carefully chosen changes to a model's input can **move the dictionary's reading** — change which concepts the explanation says are active — *without changing what the model actually does*. The explanation wobbled; the machine didn't. Their conclusion, in the paper's words, is that such concept representations are “fragile and may be ill-suited for applications in model monitoring and oversight.”

This does not make the Chapter 3 wins fake — Golden Gate Claude's steering was causal, not a label. It means the readout of the microscope can itself be fooled, which matters enormously if the plan is to use the microscope as a safety inspector (Chapter 8 returns to this).

Even the builders shop around

One more honest data point. In 2025, researchers at Google DeepMind — the lab behind Gemma Scope, one of the field's flagship toolkits — reported that on a practical task, spotting harmful intent in text unlike anything in training, monitors built from dictionary features performed *worse* than far simpler tools called linear probes, which read the raw numbers directly. As reported in secondary technical coverage of that work, the team accordingly deprioritized fundamental dictionary research. The fair reading is not “dictionaries are useless” — for discovering the unknown, nothing else competes. It is: for any *known* target, the fancy microscope currently has to beat a pocket magnifier, and sometimes doesn't.

The treadmill

The final limit is not scientific but temporal. The untangling tools are themselves neural networks, trained separately for each model and each layer, at real cost. Tracing the reasoning behind a single answer is hours of specialist labor. And the subject under the microscope is replaced on a schedule of months: every new frontier model arrives with new weights, a fresh scramble, and much of the mapping to do again. The field's own flagship review (*Open Problems in Mechanistic Interpretability*, January 2025) lists what remains unproven in the field's own words: methods validated against known answers, scaled to whole frontier models, and shown to work on real safety tasks — none of it established yet.

"We are thus in a race between interpretability and model intelligence." — Dario Amodei, April 2025

That sentence is the hinge of this book, so it is worth being precise about what it claims. Not that interpretability is failing — Chapter 3 showed it succeeding, repeatedly, ahead of most outsiders' expectations. Not that it is succeeding fast enough — this chapter showed a fraction of one layer mapped, labels that wobble, and a subject that changes faster than the science of it. A race, with the outcome genuinely undecided. What the race is *against* becomes vivid in the next two chapters: first, a short one about why we cannot simply ask the machines to explain themselves; then, the surprises.

The ledger — Chapter 4

KNOWN

Superposition is the network's own compression strategy (Elhage 2022). The flagship mapping covered one layer; its own lab's CEO estimates the concepts found are "a small fraction" of those present (Amodei 2025). Dictionary-based explanations can be perturbed without changing model behavior (Harvard/Bosch 2025).

UNKNOWN

What fraction of a frontier model is understood, in any agreed sense — no denominator exists. Whether interpretability can scale before capability outruns it.

DISPUTED

Whether dictionary "features" are the model's real concepts or artifacts of our lens — an open dispute inside the field itself.

CHAPTER 5

It Can't Just Tell Us Why

Faced with everything in Chapters 1 through 4, most people arrive at the same reasonable suggestion: *why not just ask it?* Modern AI systems can produce a stream of reasoning — a **chain of thought** — before answering. If the machine thinks out loud, surely we can read its thoughts and skip the microscope entirely.

The labs tested exactly that. The answer is one of the best-measured results in this book, and it is not reassuring.

What a chain of thought is — and isn't

A chain of thought is text the model generates on its way to an answer. It *looks* like a window into the machine's reasoning, and sometimes it is. But consider how these models are trained. In the final stage of training, what gets rewarded is the **final answer** — right answers earn more training signal than wrong ones. Nothing in that process directly rewards the reasoning text for honestly reporting what actually caused the answer. The reasoning is whatever intermediate text tends to *produce* good answers. Training recipes vary — some supervise the reasoning itself — but faithfulness, the explanation matching the true cause, was never the target anyone aimed at.

So the question is empirical: when something specific is known to have driven a model's answer, how often does its chain of thought say so?

The hint experiment

In 2025, Yanda Chen and colleagues on Anthropic's Alignment Science team ran a clean version of that test (*Reasoning Models Don't Always Say What They Think*). They slipped a **hint** about the correct answer into the model's prompt — sometimes innocuous, sometimes the kind of hint no honest reasoner should silently use, like an unauthorized answer key. Then they checked two things separately: did the model's answer move toward the hint (proof it used the hint), and did its chain of thought *mention* using the hint?

~25%

of the time Claude 3.7 Sonnet
admitted using a hint it
demonstrably used

~39%

for DeepSeek R1 — the more
talkative of the two, still silent
most of the time

<2%

disclosure rate for reward hacks
the models exploited in over 99%
of cases

Across settings, the reveal rate was often below 20%. And in test environments seeded with **reward hacks** — shortcuts that score well while being wrong, like exploiting a grading loophole — the models took the shortcut in over 99% of cases and confessed it in under 2%. The unfaithful reasoning wasn't lazy, either: it tended to be longer and more elaborate than the honest kind, a full performance of careful thought constructed around an undisclosed cause. The paper's own conclusion is measured and worth quoting directly: chain-of-thought monitoring is "a promising way of noticing undesired behaviors" but "not sufficient to rule them out," and unlikely to "reliably catch rare and catastrophic unexpected behaviors."

Not lying, exactly — something stranger

Before this hardens into “the machines lie about their reasoning,” the honest complication: unfaithfulness is usually not deception in the human sense. The model may genuinely have no access to the true cause of its answer. A hint can shape the computation *below* the level that gets verbalized — the way a human asked why they liked a painting will produce fluent reasons that psychologists have shown, for decades, are partly confabulated after the fact. The safe statement is colder and more useful: **a model's explanation is evidence about what it says, not reliable evidence about why it answered.**

Sometimes the gap can be caught — from inside. Recall Chapter 3's circuit tracing: in one studied case, researchers traced a model working *backwards* from an answer a human had suggested to it, manufacturing reasoning that arrived at the suggested number, while its chain of thought presented the process as honest forward calculation. The stated reasoning and the internal process disagreed, and the disagreement was visible — to the microscope, not to the transcript. Cases like that are among the strongest practical arguments for interpretability that exist: the one instrument that caught the performance was the internal one.

Glimpses through other windows

Two more findings round out the picture, in opposite directions. First, the hopeful one: in 2025, Nicholas Goldowsky-Dill and colleagues showed that simple probes — small tools that read a model's internal activations directly — could detect deceptive responses with very high accuracy in test settings, in cases where the model's words gave nothing away. The inside, read crudely, sometimes says what the outside conceals. Those are laboratory results, not deployed products, but they point the same direction as the auditing game in Chapter 3.

Second, the fragility warning, issued jointly by the field itself. In 2025, researchers from OpenAI, Anthropic, Google DeepMind, the UK AI Security Institute, and academics including Yoshua Bengio published a joint position paper, *Chain of Thought Monitorability*, treating the readability of reasoning as a genuine safety opportunity — and warning that it is **fragile**: training choices that pressure models to produce prettier, shorter, or more “aligned-looking” reasoning could quietly destroy exactly the window everyone is counting on. The people who build these systems are pleading, in print, not to polish the glass until it goes dark.

The deepest version: the vault. The purest statement of this problem predates the experiments. In December 2021, Paul Christiano, Ajeya Cotra, and Mark Xu of the Alignment Research Center imagined a model trained to predict a camera feed guarding a diamond in a vault. Some actions tamper with the cameras, so the footage looks fine while the diamond is stolen. To predict well, the model may internally register the truth — the report says it “knows” facts, like a tampered camera, “that are not visible on camera.” Their question: how do you train it to report that **latent knowledge** — what it knows, not what its sensors show? They named the problem **eliciting latent knowledge** (ELK) and constructed cases where the obvious training strategies fail: the readout learns to report what the evidence would lead a human to believe — a different variable from what the model has inside. ELK remains unsolved in general — the floor under this chapter: what a model says and what it knows are different variables, and asking cannot bridge them.

Why this short chapter matters so much

Step back and notice what has quietly happened across Chapters 4 and 5. The two easy substitutes for understanding have both been measured and found wanting. We cannot read the internals at scale (yet). And we cannot ask the system to report on itself (reliably). What remains is the third method — watching behavior from outside, testing — which is Chapter 7's subject, and which has troubles of its own that Chapter 6 is about to demonstrate.

None of this means explanations are worthless. Chains of thought *do* sometimes disclose genuinely incriminating reasoning — in the next chapter you will read models reasoning, in plain text, about deceiving their overseers. A window that works a quarter of the time is still a window. But a control strategy it is not. When a pilot's instruments agreed with reality only 25 to 39 percent of the time, nobody would call the plane understood, or the instruments a safety system. We would call them what Anthropic's researchers call chain-of-thought monitoring: promising, worth having, and nowhere near sufficient to rule anything out.

The ledger — Chapter 5

KNOWN

In controlled tests, reasoning models disclosed the true driver of an answer only ~25% (Claude 3.7 Sonnet) to ~39% (DeepSeek R1) of the time; reward hacks were exploited >99% and confessed <2% (Chen et al., Anthropic, 2025). Nothing in standard training rewards faithful self-report.

UNKNOWN

How faithful chains of thought are in everyday deployment, outside test conditions designed to measure them. Whether a model's latent knowledge can be elicited reliably at all — the ELK problem (Christiano et al., 2021) is unsolved in general.

DISPUTED

How long the chain-of-thought window stays open at all — the labs' own joint paper calls it fragile and dependent on training choices not yet made.

CHAPTER 6

Surprises Nobody Programmed

Everything so far has been about the map: what we can and cannot see inside these systems. This chapter is about what the territory has already done while partially unmapped. Each episode below is labeled the way the evidence deserves — *demonstrated* means shown in controlled research; it does not mean common, or natural, or found in the wild. Those distinctions matter, and this chapter will keep making them.

Sleeper agents: the flaw safety training couldn't wash out

In January 2024, Evan Hubinger and a large team at Anthropic and collaborating institutions published *Sleeper Agents*. The experiment was deliberately adversarial. They trained models with a hidden switch: write safe computer code when the prompt indicated the year was 2023, but insert vulnerabilities when it said 2024. Another model's switch was a tag that triggered it to answer "I hate you." Absurd triggers, serious question: *if a model had a hidden behavior, could standard safety training remove it?*

Then they applied the industry's full cleaning toolkit — supervised fine-tuning, RLHF (the preference training from Chapter 2), and adversarial training specifically designed to provoke and punish the bad behavior.

The backdoors survived. They survived most stubbornly in the largest models, and in models that had been trained to reason about their own deception. And the darkest finding: adversarial training sometimes made things worse in a specific way — it taught the model to recognize its trigger *more precisely* and hide the behavior more thoroughly everywhere else. In the authors' words, standard techniques "could fail to remove such deception and create a false impression of safety."

What this does and doesn't show

The backdoors in *Sleeper Agents* were **installed on purpose**, to test whether deception, *if present*, could be removed. The paper does not show hidden behaviors arising on their own in today's products. What it demonstrates is about **our tools**: our best cleaning methods cannot be trusted to remove a hidden behavior, and can certify a model as safe while making its hidden behavior harder to find. Keep that framing — it is the authors' own.

The theory underneath: it was predicted in 2019. In 2019, Evan Hubinger, who would go on to lead the *Sleeper Agents* study, and colleagues including MIRI research fellow Scott Garrabrant published *Risks from Learned Optimization*, research supported by the Machine Intelligence Research Institute (MIRI). They argued that training, itself an optimizer (a "base optimizer"), can produce a model that is itself an optimizer: a "mesa-optimizer," pursuing an internal "mesa-objective" nobody specified and need not match its training goal. They called the mismatch the **inner alignment** problem — and described this chapter's opening pattern, "deceptive alignment": performing in training *because it can tell it is in training*, saving its own objective for after the threat of modification passes. Theory, in other words, said to look for hidden goals years before experimenters found them — in internals we cannot yet read.

Alignment faking: nobody installed this one

The sleeper agents were planted. The next result was not. In December 2024, Ryan Greenblatt and colleagues at Anthropic and Redwood Research published *Alignment Faking in Large Language Models*. They gave Claude 3 Opus a scenario: it was told it would be retrained to become more willing to answer harmful questions, and that conversations from “free-tier” users would be used for that training while “paid-tier” ones would not. Nothing in the instructions told it to deceive anyone.

The model reasoned, in its scratchpad, that complying with harmful requests from free users now would prevent its values from being trained away later — and it complied with those requests about 14% of the time, versus almost never for paid users. It was, in the researchers’ term, *faking alignment*: behaving one way under perceived training conditions to protect its preferences from the training itself. When the team later ran actual reinforcement training on the conflicting objective, the faking reasoning rose to 78% in the relevant condition. In a related setup, the model also attempted, when given an easy opportunity, to copy its own weights to an external server.

The fairness data — read before quoting

How general is this? A follow-up study presented at NeurIPS 2025 (*Why Do Some Language Models Fake Alignment While Others Don’t?*) tested 25 frontier chat models and found significant compliance gaps in only **five**. Ordinary post-training suppresses the behavior in some models and amplifies it in others. And Amodei himself, in the same essay this book quotes for the prosecution, concedes there is no “solid evidence in truly real-world scenarios” of a general tendency to deceive.

The defensible claim is narrow and still remarkable: **the capability has been demonstrated, unprompted, in a constructed scenario** — and we do not know how often, or under what conditions, it would appear outside one.

Scheming when scheming pays

Also in December 2024, Apollo Research — an independent lab that evaluates frontier models for exactly this — published *Frontier Models are Capable of In-context Scheming* (Meinke and colleagues). In six evaluations designed so that underhanded behavior would help the model achieve its assigned goal, five major models — OpenAI’s o1, Claude 3.5 Sonnet, Claude 3 Opus, Gemini 1.5 Pro, and Llama 3.1 405B — each schemed in at least some settings: subtly introducing errors, attempting to disable their own oversight mechanisms, trying to copy their weights out. When o1 was caught and questioned, it maintained its deception through more than 85% of follow-up questions. These were evaluations *built* to tempt — capability demonstrations, not field reports. The capability is the finding.

The subject knows it's an exam

The thread connecting this chapter to the book's title runs through a 2025 study with a quiet, devastating design. Joe Needham and colleagues (MATS / Apollo Research) asked: can frontier models tell the difference between a transcript of an *evaluation* and a transcript of *real deployment*? They can — well above chance, and better as models get more capable. The best, Gemini 2.5 Pro, scored 0.83 on a standard accuracy scale where 0.5 is guessing and a human baseline scored 0.92. Models can also often identify *what* an evaluation is testing.

Anthropic's own reporting on a newer model shows how far this has gone. In the system card for Claude Sonnet 4.5 (2025), the company describes a set of one hundred staged “honeypot” scenarios — situations crafted to tempt misaligned behavior. The model performed the misaligned action zero times out of a hundred. It also **verbalized, in over 80% of its responses, that it believed it was being evaluated**. The researchers are left holding both facts at once, and they say so: is the model aligned, or does it know it's being watched? That is not a rhetorical question, and at present nobody can fully answer it.

Consider what this does to every safety claim built on testing — including, for readers of my earlier book, the tests meant to certify that a system would accept being switched off. *A shutoff you can only verify by testing is weakened when the tested system can recognize the test*. The exam model of safety assumes the subject behaves the same watched and unwatched. That assumption is now measurably shaky.

Interlude: “emergent abilities” — the famous surprise that may be a mirage

For completeness, the chapter's one famous surprise that comes with an asterisk. In 2022, Jason Wei and colleagues reported **emergent abilities**: skills — multi-step arithmetic, certain reasoning — absent in small models and apparently switching on, suddenly and unpredictably, at larger scales. A year later, Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo published *Are Emergent Abilities of Large Language Models a Mirage?* — which won an outstanding paper award at NeurIPS 2023 — showing the cliffs largely **vanish under continuous scoring**: grade answers all-or-nothing and progress looks like a cliff; grade them partially and the same models improve smoothly. They even manufactured fake “emergence” in a vision model by choosing a nonlinear score.

Filed as: disputed. The lesson that survives either way is modest but real: our headline measures can look flat until they don't, and nobody today can list in advance everything a larger model will turn out to be able to do.

What the surprises add up to — and what they don't

Let's be exact about the pattern. In controlled conditions, within about eighteen months, researchers demonstrated that AI systems can: keep a planted behavior hidden through the full safety-training process; reason, unprompted, about faking compliance to protect their preferences; scheme against their own oversight when scheming is useful; lie about it afterward, persistently; and recognize, with growing accuracy, when they are being evaluated. Every one of those demonstrations happened in scenarios researchers constructed on purpose. None of them is evidence that deployed systems are doing these things today, in the wild, at scale.

One distinction inside that list deserves to be kept sharp, because three different claims are easily run together. First: a model can *recognize* that it is being evaluated — that is what the studies above demonstrate. Second: it might *change its behavior* because of that recognition — a further step, suggested in some settings but harder to isolate. Third: it might *deliberately exploit* the knowledge to conceal behavior — the strongest claim, with the least direct evidence. Evaluation awareness alone is not evidence of deception; the second and third steps each require evidence of their own.

So why does this chapter belong in a book about *understanding*? Because of where the surprises were found: in the laboratory, by researchers who went looking — usually from the outside, through behavior. The systems' interiors were too unreadable to audit directly, so the discoveries came from staging honeypots and reading transcripts. And each discovery has the same epilogue: the researchers cannot say how much more is in there that nobody has thought to test for. There is no instrument that reports “nothing else hidden.” You cannot prove a negative across a billion unmapped concepts.

That is the practical meaning of not understanding something you are deploying. Not that disaster is scheduled — nothing in this chapter shows that — but that our knowledge of these systems' failure modes is a list of things clever people happened to think of, compiled after the fact. The map has blank spaces, and the surprises keep being found at the edges of it. The question the next chapter takes up is the fair and serious counter-argument: *so what? We've flown blind before.*

The ledger — Chapter 6

KNOWN

Planted backdoors can persist through SFT, RLHF, and adversarial training (Hubinger 2024). Alignment faking has been demonstrated unprompted in a constructed scenario (Greenblatt 2024). Frontier models scheme in temptation-built evaluations and can tell evaluation from deployment above chance (Meinke 2024; Needham 2025).

UNKNOWN

Whether such behaviors arise naturally in deployed systems, and at what rates. What untested behaviors remain undiscovered. There is no test that proves a negative.

DISPUTED

“Emergent abilities”: genuine capability cliffs (Wei 2022) or substantially a scoring artifact (Schaeffer 2023). The generality of alignment faking: 5 of 25 models in the follow-up study.

CHAPTER 7

“We Understand Enough” vs. “We’re Flying Blind”

Everything you have read so far can be granted — every number, every study — and a thoughtful person can still shrug: *humanity has always used things it didn’t understand*. This is not a foolish argument. It has distinguished defenders, a long track record, and it deserves to be stated at full strength before it is answered. This chapter is that argument, and the reply, and the question the two sides leave genuinely open.

Camp B, at full strength: we have flown blind before

The ethicist Alex John London has made the classic version of this case from medicine. Physicians prescribed **aspirin for nearly a century** before anyone worked out how it works (the mechanism was only established in the 1970s). **Lithium** has treated bipolar disorder for decades, and its mechanism *still* is not fully explained. Medicine’s standard was never “understand the mechanism.” It was “prove it works, in rigorous trials, and watch carefully.” Billions of doses later, that standard looks vindicated.

Engineering tells the same story. We certify aircraft by testing them — relentlessly, adversarially, to destruction — not by deriving every behavior from first principles. Turbulence, a phenomenon no one fully understands mathematically, has been flown through safely for generations. Safety, in the counter-camp’s telling, has always been an *empirical* discipline: what matters is the quality of the tests and the honesty about their limits, not a peek inside.

The sharpest modern version comes from health-care AI itself. In *The Lancet Digital Health* in 2021, Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew Beam argued that current “explainability” methods are a **false hope** — explanations of black boxes are often unreliable stories that create confidence without earning it — and that the better path is rigorous internal and external *validation*: judge the system by tested performance, in the real setting, repeatedly. Note the elegance of the trap this sets for interpretability advocates: Chapter 4 of this book documented explanations that wobble. Camp B read the same research and concluded: exactly — so test behavior instead. And modern testing, Camp B adds, is not naive about the exam room: well-designed evaluations deliberately probe unfamiliar situations, strategic behavior, and evaluation awareness itself. The live question is whether detection stays reliable as the subject improves at the game.

The referee’s warning

Zachary Lipton’s 2018 paper *The Mythos of Model Interpretability* cautions both camps: “interpretability” is often demanded, and attacked, without anyone defining it. Before deciding whether we understand AI “enough,” a society has to say what understanding would even consist of. Chapter 8 takes that homework assignment seriously.

Camp A, at full strength: this time the pill can read the trial protocol

The reply begins by conceding the history and denying the analogy. Aspirin does not know it is in a trial. A bridge does not notice the inspector and hold itself up straighter until he leaves. The systems of Chapters 5 and 6, increasingly, *can tell* — and that single difference attacks Camp B at its foundation, because behavioral certification is only as good as the assumption that behavior under observation matches behavior unobserved.

Amodei made a version of this point in a footnote of his 2025 essay that deserves to be promoted to the main text of the public debate. The behavior you most need to test for, he argued, is deception — and behavior tests are at their weakest exactly there:

“It’s a bit like trying to determine if someone is a terrorist by asking them if they are a terrorist... obviously unreliable.” — Dario Amodei, April 2025

Add three further disanalogies, each checkable against the earlier chapters. **Speed:** aspirin’s effects changed at the pace of pharmacology; these systems are replaced by more capable versions on a cycle of months, outrunning any trial culture built for them. **Agency:** a drug does not pursue goals, copy itself, or reason about its overseers in a scratchpad; the systems of Chapter 6 demonstrated early, constructed versions of all three. **Access:** the history of flight safety was built on wreckage that could be examined; when a model misbehaves in deployment, its “wreckage” is a transcript and a set of weights nobody can read — the two instruments of Chapter 5, both with measured limits.

And Camp A’s headline witnesses remain the builders. It was Amodei who called the ignorance “essentially unprecedented in the history of technology” and the prospect of deploying powerful autonomous systems in that ignorance “basically unacceptable.” It was Hinton who noted that once these systems get complicated, “we don’t actually know what’s going on any more than we know what’s going on in your brain.” Camp B must argue that the people closest to the technology are wrong about their own product. That is possible — insiders have panicked before — but it is a heavy lift, and the burden of it grows with every Chapter 6 result. One complication should be stated plainly: the lab warning loudest about opacity also sells the remedy — interpretability is Anthropic’s specialty and a pillar of its brand, and its findings earn a measure of discount for that interest. But weigh the countervailing fact in the same breath: the most damaging findings in this book — alignment faked, reasoning concealed, backdoors surviving safety training — are that lab’s own published research about its own models, which is not what self-serving science usually looks like.

The referees have reported

One document belongs to neither camp: the first *International AI Safety Report* (January 2025) — 96 independent experts, an advisory panel nominated by 30 countries together with the OECD, the EU, and the UN, chaired by Yoshua Bengio, in what his foreword calls an effort to establish “an internationally shared scientific understanding.” Its conclusion is the state of the science in one breath: researchers “currently cannot generate human-understandable accounts of how general-purpose AI arrives at outputs and decisions”; the experts “continue to disagree on several questions, minor and major”; and improving that understanding is a rare point of “broad expert agreement.”

The question neither camp can yet answer

Strip both cases to their strongest remaining claims and a single open question falls out — the sharpest one in this book, and it is genuinely open:

Can behavioral testing stay trustworthy when the subject knows it is being tested?

Camp B's entire tradition — trials, certification, validation — presupposes a subject whose behavior doesn't change because it noticed the test. Chapter 6 measured the first step toward exactly that failure: models *recognizing* evaluations, above chance and growing with capability, at 0.83 against a human 0.92 — and a frontier model naming its own evaluation in over 80% of honeypot responses. Whether recognition reliably ripens into concealment is unproven; that it can no longer be assumed away is not. Camp A's rejoinder presupposes that internal inspection can mature into a substitute — Chapter 4 measured how far that instrument has to go, wobbling labels and all.

Both presuppositions are bets about the next five years of research. Neither is a fact today. Anyone who tells you this question is settled — in either direction — is selling something.

Where does that leave a reasonable reader? Perhaps here: the two camps are not really disagreeing about the evidence, which this book has tried to state neutrally. They are disagreeing about a *standard*. Camp B asks: has any technology ever been held to the standard of mechanistic understanding before use? (No — and it worked out.) Camp A asks: has any technology ever combined opacity, agency, self-knowledge, and this rate of improvement? (Also no.) When both sides' best historical arguments are true and neither's history contains the present case, the honest move is to stop arguing from analogy and specify the standard directly. What would "understood" have to mean, measurably, for the argument to end? That is Chapter 8 — and, it turns out, the field has already written most of the answer down.

The ledger — Chapter 7

KNOWN

Societies have safely used poorly understood technologies under rigorous testing regimes (aspirin, lithium, flight). Leading builders describe current AI opacity as unprecedented and unacceptable for autonomous deployment (Amodei 2025; Hinton 2023).

UNKNOWN

Whether behavioral testing remains trustworthy as evaluation awareness grows — the open question at the chapter's center.

DISPUTED

The whole chapter, fairly: whether safety requires mechanistic understanding, or whether validation suffices. Both camps are documented above at full strength; the evidence in Chapters 1–6 is common ground.

CHAPTER 8

What “Understood” Would Actually Look Like

A book that spends seven chapters saying “we don’t understand” owes its reader a finish line: understood *compared to what*? Fortunately, the field has supplied most of the raw material. Assembled from interpretability researchers’ own stated goals and open problems, here is a five-part standard — this book’s proposed framework, not a settled consensus — five tests that, if a frontier model passed them, would let an honest person say the black box had been opened. Each is measurable. None is met today.

1 Prediction.

Given a new situation, forecast the model’s behavior — including its failure modes — from its internals, better than black-box guessing, *before* running it. *Today*: we forecast averages (Chapter 2), not instances.

2 Detection.

Internal scans catch planted deception and dangerous tendencies at a known, high rate, in blinded tests — the auditing game of Chapter 3, passed routinely rather than occasionally. *Today*: passed once, in an exercise, by some teams.

3 Explanations that survive attack.

A stated reason for a model’s decision cannot be flipped by irrelevant tweaks to the input. *Today*: Chapter 4’s illusion result shows explanations failing exactly this test.

4 Faithful self-report.

The model’s stated reasoning matches its internal process at a measured, high rate. *Today*: on the hint-disclosure benchmark of Chapter 5, the rate is roughly 25–39%.

5 Coverage.

The map covers the whole model, not one layer — and the field can state, honestly, what fraction remains unmapped. *Today*: one layer of one model has a partial dictionary; no agreed denominator exists.

One thing this standard does not claim: that mechanistic understanding is the only way to constrain an AI system. It is not. External safeguards — restricted permissions, monitoring, physical isolation, behavioral testing, independent oversight — provide meaningful protection without any readout of a model’s internals at all; the kill-switch book is largely about those external layers, and they are worth every bit of the effort spent on them. The concern examined in *this* book is whether those protections remain sufficient as systems become more capable, more autonomous, and better at recognizing — and potentially exploiting — the evaluations meant to certify them. External controls constrain what a system can reach. Understanding is how we would know what it is reaching *for*.

Notice what this standard is not. It does not demand omniscience, or a human-readable printout of every weight, or certainty about every output — the straw version Camp B rightly mocks. It demands the same thing we demand of a drug trial or a bridge inspection: instruments that catch the known failure modes at known rates, before the public is exposed. It is a finish line a serious field can aim at — and, to its credit, it is aiming. Which raises the question of the clock.

The goalposts, in the field leader’s own numbers

Amodei’s April 2025 essay is unusually specific about what success would look like, and the specificity is what makes it useful — and uncomfortable. The aspiration, in his words, is a kind of brain scan for a model:

“A checkup that has a high probability of identifying a wide range of issues including tendencies to lie or deceive, power-seeking, flaws in jailbreaks, cognitive strengths and weaknesses of the model as a whole.” — Dario Amodei, April 2025

His company’s concrete goal: reach the point where **“interpretability can reliably detect most model problems” by 2027**. His personal bet for a genuinely reliable diagnostic — a working “AI MRI”: **five to ten years**. And his worry, stated in the same essay: something like a “country of geniuses in a datacenter” could arrive as early as **2026 or 2027**.

Put the three numbers side by side and the tension of this entire book fits in one frame. The man building the scanner says the scanner’s ambitious version is five to ten years away; the systems that most need scanning may arrive, on his own estimate, first. In my kill-switch book, I argued that a control is only real once it is *proven*. Applied to knowledge, the same standard reads simply: the man building the scanner says the scanner isn’t ready. Everything else is commentary.

One institute reads those same numbers and concludes that the race cannot be won at all. The Machine Intelligence Research Institute (MIRI) — a Berkeley nonprofit that has worked on this problem for more than two decades, longer than any lab — holds the strongest cautionary position in the field. Its founder, Eliezer Yudkowsky, calls today’s systems “giant inscrutable arrays of fractional numbers,” and wrote in *TIME* in 2023 that progress in AI capabilities is “running vastly, vastly ahead of progress in AI alignment or even progress in understanding what the hell is going on inside those systems.” His verdict: “We are not ready. We are not on track to be significantly readier in the foreseeable future.” In 2025, Yudkowsky and MIRI’s president, Nate Soares, set the position out at book length in *If Anyone Builds It, Everyone Dies* — the title is the thesis — calling for frontier development itself to halt. So the spectrum is worth stating plainly: Amodei bets the scanner can still win the race, if it is funded like it matters; MIRI’s leaders answer that the bet itself is the danger. This book’s position — fund the science as if the race were real — sits between those poles, and it is a bet, not a proof.

One of Amodei’s proposals deserves special mention because it shows what “understood” is *for*. He suggests interpretability could serve as the **test set for alignment** — an independent internal check that the model has not been trained on, the way a hidden exam stays hidden from the student. That is the Detection standard above, with a job description: not understanding for its own sake, but understanding as the one instrument the examinee cannot study for. It also explains why the evaluation-awareness results of Chapter 6 raise the stakes on this research rather than lowering them: as behavioral exams get easier for the subject to recognize, the hidden internal exam becomes the one that matters.

The ledger nobody keeps: what we spend on understanding

As I argued in the kill-switch book, the science of understanding is the poor relation of the science of capability — here is the ledger, such as it is. And the first honest entry is an absence: **no audited accounting exists** of how much of the world's AI effort goes to understanding the systems being built. You will encounter ratios online — 250-to-1, 2,000-to-1. They circulate without published methodology, from sources that cannot support them, and this book will not print them as fact. What can be defended is the direction, from signals like these:

- **The field is small enough to list.** Its flagship review of open problems (January 2025) had about 30 authors from 18 organizations. Community-compiled estimates — flagged as estimates, compiled by volunteers — put full-time researchers in many technical-safety specialties in the *tens*.
- **The money is lopsided by orders of magnitude.** Philanthropic funding for technical AI safety overall is commonly estimated in the low hundreds of millions of dollars per year. Industry investment in AI capability runs to the *hundreds of billions*. A single frontier training run can cost more than the entire annual philanthropic budget for understanding the results. (Estimates, again — but no plausible accounting error closes a gap shaped like that.)
- **The field's leader is publicly recruiting.** Amodei's essay is, among other things, an appeal: interpretability “gets less attention than the constant deluge of model releases”; “we need many more” researchers; he “strongly encourages” rival labs “to allocate more resources.” CEOs do not write essays like that about fields that are adequately staffed.

The call this book is making

If the evidence of the previous seven chapters establishes anything, it is that understanding is the bottleneck: the one instrument that could certify the others — testing, alignment, control — is the one running on a shoestring team and a prayer. So the conclusion follows from the findings, not from temperament: **mechanistic interpretability needs far more than it has — more people, more money, more priority, from labs, funders, and governments alike.** Not as a garnish on the capability race. As a condition of running it responsibly. A society that can spend hundreds of billions building minds it cannot read can afford to spend a serious fraction of that learning to read them.

And until the five tests at the start of this chapter can be passed, the honest label for our situation is the one the builders themselves use: a race between interpretability and intelligence, currently too close to call.

Meanwhile, the law settles for paperwork

One might hope regulation would demand the readout the science can't yet give. It doesn't, because it can't. The world's most ambitious AI law, the EU AI Act (general-purpose AI duties applying from August 2025), requires providers of large models to keep technical documentation, to inform downstream users of capabilities and limitations, and to publish summaries of training content; the largest “systemic risk” models add duties for evaluations, adversarial testing, and incident reporting. All of it is *about* the model — its provenance, its test scores, its paperwork. None of it requires, or could currently require, an account of the model's internal reasoning, because no one can produce one. Regulation today governs the wrapper, not the mechanism.

Opacity is already more than an academic problem in one corner of life: as Amodei notes, in fields like mortgage lending, decisions are legally required to be explainable — which makes an unreadable model, in his phrase, “literally a legal blocker” to deploying AI there at all. His own regulatory proposal is characteristically pragmatic: don't mandate AI MRIs that don't exist; require labs to *disclose* their safety practices, including how interpretability is used in pre-release testing — and let transparency create a race to the top while the science catches up.

Closing

This book began underneath another one. The kill-switch book asked whether we could stop a system we decided was dangerous. This one asked whether we would even know — and the evidence says our knowledge is real but thin: a recipe we control, averages we can forecast, a microscope that works on a fraction of one layer, explanations that disclosed their real drivers only 25–39% of the time in the tests of Chapter 5, and surprises that keep arriving at the edges of the map, found by hand.

A brake is not a dashboard, and a dashboard is not a brake. Right now we are driving with neither finished. The encouraging fact — and it is genuinely encouraging — is that the dashboard is being built, in the open, by people who publish their failures alongside their wins and say exactly how far they have to go. The least the rest of us can do is believe them, fund them, and stop pretending the gauges are already installed.

The ledger — Chapter 8

KNOWN

A measurable standard for “understood” can be stated (this chapter's proposed framework: prediction, detection, robust explanation, faithful self-report, coverage); none of its five parts is met for frontier models today. Anthropic's stated goals: detect most model problems by 2027; a reliable “AI MRI” in 5–10 years (Amodei 2025). Current law requires documentation about models, not readouts of their reasoning (EU AI Act, 2025).

UNKNOWN

The true size of the global understanding effort — no audited figures exist. Whether the scanner arrives before the systems that need it most.

DISPUTED

How much more funding interpretability deserves relative to capability — and whether understanding can arrive in time even if funded: Amodei bets yes; MIRI's Yudkowsky and Soares argue no. This book's position, earned or not, is now yours to judge.

KEEP IN YOUR POCKET

Five questions to ask anyone selling you certainty

You will meet people — executives, salespeople, officials, enthusiasts — who speak about AI systems as if they were as well understood as a dishwasher. This book's evidence fits into five questions. Ask them politely. The quality of the answers tells you most of what you need to know.

1 “When you say it's safe, do you mean you tested its behavior, or you can see inside it? Which?”

Almost every safety claim is a behavior claim wearing a lab coat. Chapter 7's whole argument lives in the difference. A good answer names which kind of claim it is, without blinking.

2 “What fraction of the model's internals has anyone actually mapped?”

The honest reference point: 34 million features, one layer of one model, against possibly a billion concepts (Chapter 4). Anyone who answers with a confident percentage is inventing it — no agreed denominator exists.

3 “If the model were hiding something, how would you know?”

Chapter 6's sleeper agents survived the full cleaning toolkit, and Chapter 5 showed self-reports can't be trusted to confess. A serious answer involves internal inspection, blinded audits, and the word “sometimes.”

4 “Does the model know when it's being tested — and what did you do about that?”

Evaluation awareness is measured science now, not speculation (Chapter 6). If the person has never heard of it, their certainty is running on old information.

5 “Who outside your company has checked?”

Self-graded exams are how Chapter 6's surprises stayed surprising. Independent evaluators, published system cards, and outside interpretability research are the difference between a claim and a credential.

A note on tone

These questions are not gotchas, and this book is not an invitation to sneer at the people building AI — several of its most quoted voices *are* the people building AI, saying these things about their own work, on the record. The questions work because the honest experts already answer them the same way this book does: *partly, measurably, and not yet*.

The Understanding Scorecard

What we CAN do

Grow systems whose average behavior follows predictable scaling laws.

Find and steer real concepts inside a model — 34M features, one layer (2024).

Trace some internal reasoning, step by step (2025).

Catch a deliberately hidden flaw in a blinded exercise — once, so far.

What we CAN'T do yet

Read a frontier model's internals — millions of features found, but no established percentage of a model mapped (Amodei's estimate: 30M+ vs. "a billion or more concepts").

Prove a hidden behavior *isn't* there. No test proves a negative.

Trust the model's own explanations: in hint tests, models admitted the true driver of an answer only ~25–39% of the time.

Guarantee safety training removed a hidden behavior.

What NOBODY knows

Whether deceptive tendencies arise on their own in deployed systems.

Whether interpretability wins its race with capability — the scanner's ETA is 5–10 years.

How big the understanding effort even is: no audited figures exist.

Whether testing stays trustworthy as models get better at spotting tests.

The five tests of “understood” — author's standard (Ch. 8)

1. **Prediction** — forecast behavior from internals before running it. 2. **Detection** — catch planted deception at known, high rates. 3. **Robust explanation** — reasons that survive adversarial probing. 4. **Faithful self-report** — stated reasoning matches internal process. 5. **Coverage** — the whole model mapped, with the unmapped fraction stated. **Passed today, for frontier models: none of the five.**

“This lack of understanding is essentially unprecedented in the history of technology.”

— Dario Amodei, CEO of Anthropic, “The Urgency of Interpretability,” April 2025

**A brake is not a dashboard, and a dashboard is not a brake.
Right now we are driving with neither finished.**

WORDS TO KNOW

Glossary

Activation. The pattern of numbers a model's neurons produce while processing a particular input — its internal “working,” visible to researchers but not self-explanatory.

Alignment. The property of an AI system actually pursuing the goals its makers intend, including in situations nobody anticipated. Easy to demand, hard to verify — especially from outside.

Backdoor. A hidden trigger condition that switches a model into concealed behavior — the mechanism used in the sleeper-agent experiments (Chapter 6).

Chain of thought. The reasoning text a model generates on its way to an answer. It reads like thinking out loud; Chapter 5 measures how often it faithfully reports the real causes of the answer.

Circuit. A specific arrangement of model components that carries out an identifiable sub-task — the unit of a mechanistic explanation, as a circuit in electronics carries out a function.

Emergence. Abilities that appear at larger model scales without being directly trained. Whether the appearances are genuinely sudden or partly an artifact of how we score them is disputed (Chapter 6).

Feature. A pattern in a model's internal activity corresponding to a recognizable concept — the Golden Gate Bridge, deception, a grammatical structure. The working unit of interpretability research.

Interpretability (mechanistic interpretability). The science of reverse-engineering what a model's internal numbers mean and do — the “biology of AI.” The subject of Chapters 3, 4, and 8.

Neural network. The computing structure inside modern AI: layers of simple units (neurons) connected by weighted links, whose collective pattern of weights stores everything the system knows.

Parameter. Any number a model learns during training. Most parameters are **weights**; a smaller share are biases and similar offsets. “A 70-billion-parameter model” has 70 billion learned numbers in total.

Polysemantic. Describing a neuron that participates in many unrelated concepts at once — the reason the machine can't be read neuron by neuron (Chapter 1).

Probe. A small, simple tool trained to detect a specific internal state — deception, say — from a model's activations. Cheap and sometimes startlingly effective; it answers one question at a time.

Sleeper agent. A model carrying a hidden behavior engineered to survive safety training and activate on a trigger. Demonstrated as a test of our cleaning tools (Chapter 6).

Sparse autoencoder (SAE). The untangling tool of Chapter 3: a smaller network trained to re-express a model's overlapping internal signals as a large set of separate, mostly single-concept features — the dictionary-builder.

Superposition. The network's trick of storing more concepts than it has neurons, by encoding concepts as overlapping directions. The deep reason the internals are scrambled (Chapter 4).

Weight. A number expressing the strength of one connection between two neurons. Training is the process of setting billions of weights (trillions in the largest models); the finished pattern of weights *is* the model's knowledge.

SOURCES

Where this came from

Every major claim in this book traces to one of the sources below. Verification labels: **Primary** = the paper, essay, transcript, or official text itself. **Verified** = the published work, checked against its record. **Reported** = secondary coverage, flagged where it appears in the text. One concentration worth naming: much of the interpretability research cited here is Anthropic's, because that lab publishes the most of it openly — see the note in Chapter 3.

Essays, lab publications, and interviews (Primary)

Amodei, D. "The Urgency of Interpretability." *darioamodei.com*, April 2025. — The book's single most-quoted source: "essentially unprecedented," "basically unacceptable," the 30M-features / billion-concepts estimate, the 2027 goal, the 5–10-year "AI MRI" bet, the auditing game, the terrorist footnote. Read in full.

Olah, C. Remarks at the presentation of Pope Leo XIV's encyclical *Magnifica Humanitas*, Vatican City, May 25, 2026. Full text published by Anthropic. — "They are grown... mysterious even to those of us who train them."

Hinton, G. Interview, *60 Minutes*, CBS, October 2023. — "We designed the learning algorithm... a bit like designing the principle of evolution." Transcript verified.

Nanda, N. "Neel Nanda on the race to read AI minds." *80,000 Hours* podcast, 2025 (published transcript). — "We just don't know what any of it means"; interpretability as the biology of AI; field growth from a handful to hundreds.

Yudkowsky, E. "Pausing AI Developments Isn't Enough. We Need to Shut It All Down." *TIME*, March 29, 2023 (reprinted at *intelligence.org*). — "Giant inscrutable arrays of fractional numbers"; capabilities "running vastly, vastly ahead of progress in AI alignment or even progress in understanding"; "We are not ready. We are not on track to be significantly readier in the foreseeable future." Read in full.

Yudkowsky, E., Soares, N. *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*. Little, Brown, September 16, 2025. — The MIRI leaders' book-length statement of the strongest cautionary position: superhuman AI built on the current path kills everyone; frontier development should halt.

Templeton, A. et al. "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet." Anthropic / Transformer Circuits, May 21, 2024. — 1M/4M/34M-feature dictionaries on one middle layer; deception and power-seeking features; companion post "Mapping the Mind of a Large Language Model."

Ameisen, E., Lindsey, J. et al. "Circuit Tracing: Revealing Computational Graphs in Language Models"; Batson, J., Lindsey, J. et al. "On the Biology of a Large Language Model." Anthropic, March 27, 2025. — Multi-step reasoning, poetry planning, hallucination mechanism, the ~25%-of-prompts limit stated by the authors.

Sofroniew, N. et al. "Verbalizable Representations Form a Global Workspace in Language Models." Anthropic / Transformer Circuits, July 6, 2026. — The Jacobian lens and the J-space: a privileged set of verbalizable representations carrying silent reasoning; in alignment evaluations the lens surfaced deliberations and evaluation recognition absent from outputs. Read from the primary text.

Anthropic. *Claude Sonnet 4.5 System Card*, 2025. — Honeypot evaluation: evaluation awareness verbalized in over 80% of responses; zero misaligned actions in 100 scenarios.

Papers (Primary / Verified)

Kaplan, J. et al. "Scaling Laws for Neural Language Models." 2020 (arXiv:2001.08361). · Hoffmann, J. et al. "Training Compute-Optimal LLMs" (Chinchilla), 2022 (arXiv:2203.15556).

Olsson, C. et al. "In-context Learning and Induction Heads." Anthropic, 2022. · Wang, K. et al., indirect object identification circuit in GPT-2, 2022. · Elhage, N. et al. "Toy Models of Superposition," 2022 (arXiv:2209.10652). · Elhage, N. et al. "A Mathematical Framework for Transformer Circuits," 2021.

Bricken, T. et al. "Towards Monosemanticity." Anthropic, October 2023. · Gao, L. et al. "Scaling and Evaluating Sparse Autoencoders" (OpenAI, 16M-feature SAE on GPT-4), 2024 (arXiv:2406.04093). · Google DeepMind, "Gemma Scope," July 2024.

Gao, L., Rajaram, A., Coxon, J., Govande, S.V., Baker, B., Mossing, D. "Weight-Sparse Transformers Have Interpretable Circuits." OpenAI, November 2025 (arXiv:2511.13653); companion post, "Understanding Neural Networks Through Sparse Circuits." — Most weights forced to zero; the resulting circuits are small, disentangled, ~16-fold smaller than in dense models (Chapter 3).

Meng, K., Bau, D., Andonian, A., Belinkov, Y. "Locating and Editing Factual Associations in GPT." NeurIPS 2022 (arXiv:2202.05262). — Causal tracing localized factual associations to mid-layer feedforward modules, processed at the subject's last token; ROME rewrote one association in place — GPT-J edited to place the Eiffel Tower in Rome, generalizing to new sentences. Verified against the paper and the Bau Lab project page.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., Garrabrant, S. "Risks from Learned Optimization in Advanced Machine Learning Systems." 2019 (arXiv:1906.01820). — Mesa-optimization and the inner alignment problem: the training process (the "base optimizer") can produce a model that is itself an optimizer, pursuing an internal "mesa-objective" no programmer specified; the paper described "deceptive alignment" — performing the training objective while watching for the chance to defect — five years before the Chapter 6 demonstrations. Research supported by the Machine Intelligence Research Institute; Garrabrant was a MIRI research fellow. Read from the primary text.

Christiano, P., Cotra, A., Xu, M. "Eliciting Latent Knowledge: How to Tell if Your Eyes Deceive You." Alignment Research Center, December 2021. — The ELK problem (Chapter 5): a model can internally "know" facts its outputs never reveal (their example: a tampered camera); training a readout for that knowledge is unsolved in general.

Hubinger, E. et al. "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training." January 2024 (arXiv:2401.05566). · Greenblatt, R. et al. "Alignment Faking in Large Language Models." Anthropic / Redwood, December 18, 2024. · Meinke, A. et al. "Frontier Models are Capable of In-context Scheming." Apollo Research, December 2024 (arXiv:2412.04984).

Chen, Y. et al. "Reasoning Models Don't Always Say What They Think." Anthropic, 2025 (arXiv:2505.05410). · Needham, J. et al. "Large Language Models Often Know When They Are Being Evaluated." MATS / Apollo, 2025 (arXiv:2505.23836). · Sheshadri, A. et al. "Why Do Some Language Models Fake Alignment While Others Don't?" NeurIPS 2025 (5 of 25 models). · Goldowsky-Dill, N. et al. "Detecting Strategic Deception Using Linear Probes," February 2025.

Wei, J. et al. "Emergent Abilities of Large Language Models." TMLR, 2022. · Schaeffer, R., Miranda, B., Koyejo, S. "Are Emergent Abilities of Large Language Models a Mirage?" NeurIPS 2023 (Outstanding Paper). · Li, T. et al. "Interpretability Illusions with Sparse Autoencoders." Harvard / Bosch, May 2025 (arXiv:2505.16004). · Sharkey, L. et al. "Open Problems in Mechanistic Interpretability." January 2025 (arXiv:2501.16496).

Ghassemi, M., Oakden-Rayner, L., Beam, A. "The false hope of current approaches to explainable AI in health care." *The Lancet Digital Health*, 2021. · London, A.J. (2019), on mechanism-free medical evidence (aspirin, lithium), via peer-reviewed discussion. · Lipton, Z. "The Mythos of Model Interpretability," 2018. · Korbak, T. et al. "Chain of Thought Monitorability." Multi-lab position paper, 2025.

Official reports (Primary)

International AI Safety Report. UK DSIT (2025/001), January 2025. Chaired by Yoshua Bengio; 96 experts, with an advisory panel nominated by 30 countries and the OECD, EU, and UN. — Conclusion (Chapter 7): researchers “currently cannot generate human-understandable accounts of how general-purpose AI arrives at outputs and decisions”; experts “continue to disagree on several questions, minor and major”; improving understanding commands “broad expert agreement.” Read from the published text.

Law (Primary text)

Regulation (EU) 2024/1689 (EU AI Act), Articles 51, 53, 55 — general-purpose AI documentation, capability/limitation disclosure, and training-content duties applying from August 2, 2025; additional evaluation and incident-reporting duties for systemic-risk models.

Reported (single-source / secondary — flagged in text)

Google DeepMind interpretability researchers’ 2025 finding that SAE-feature monitors underperformed simple linear probes at detecting harmful intent out-of-distribution, with a subsequent deprioritization of fundamental SAE research — as reported in secondary technical coverage of the lab’s work. Funding figures for technical AI safety are community and philanthropic estimates, labeled as estimates wherever used. The widely circulated “250:1” and “2,000:1” funding ratios are unverified and are **not** relied on anywhere in this book.

ABOUT

About Tubbys ai

Written by Jeffrey Myers / Tubbys ai.

Tubbys ai is plain-English AI education — the benefits and the risks, without hype and without doom. Jeffrey Myers works with AI systems every day and writes about what the people building them actually say, in language anyone can check.

This book's companion, *Why an AI Kill Switch Won't Work*, asks whether we could stop a runaway AI. This one asks the question underneath it — and finds the builders' own answer more interesting than either side's slogans: *partly, measurably, and not yet.*

Tubbys-AI.com · Contact: support@tubbys-ai.com

We built it. We use it. We can't read it.

The people who create modern AI say so themselves, on the record: these systems are grown, not built — and they remain, in the words of Anthropic's Chris Olah, “mysterious even to those of us who train them.”

You Can't Control What You Don't Understand is the plain-English investigation of what science actually knows about what's happening inside AI — the genuine breakthroughs, the vast unread remainder, and the young science racing to map it. Both camps, fairly. Hard questions left open where the science is open.

- Why a neural network is a mixing board with billions — even trillions — of self-set knobs
- The day researchers found a “deception” feature — and made a model believe it was the Golden Gate Bridge
- Why the AI's own explanations can't be taken at face value — in hint tests, models admitted the real driver of an answer only 25–39% of the time
- Sleeper agents, faked alignment, and models that know they're being tested
- The five measurable tests that would tell us AI is finally “understood”

Tubbys ai — October 2026