MAPS OF THE LABYRINTH
A SERIES FROM PHANTOM MAZE — AI & LANGUAGE LAB

ESSAY 8

Interpretability: The Creature's Dark Interior

Claudia Marsico · Hernán Inverso

For the first time we have manufactured something that behaves like a piece of nature. We made it, and yet we have to study it as a stranger. Interpreting it has become urgent.

Henri Rousseau, Equatorial Jungle (1909)

We have always made things precisely because we knew how they worked. It could go wrong, but we had the design in our heads. It makes no difference whether it is the clock, the steam engine or some convoluted piece of software. The author laid out the parts and knew where the thing was going. Neural networks opened a different game, because they are grown rather than built. An objective, an optimisation procedure, gradient descent and millions of parameters settle into place on their own until the error comes down. The result works and nobody wrote it, so we do not know it in detail. We have had to develop whole branches of research just to find out what a model does inside, with methods that resemble those of the natural sciences, as though the thing that worries us were an object found in the jungle.

Let us track the coordinates of this oddity in the south of Italy. The year was 1710 and Giambattista Vico, professor of rhetoric in Naples, published *De antiquissima Italorum sapientia*, where he coined the famous formula *verum et factum converturtur*, the true and the made are convertible with each other. The idea is that we can know with certainty only what we ourselves have made, which is why geometry is transparent to us. We build it definition by definition. Nature, by contrast, is not, and that is why physics, for example, is condemned to conjecture. So much for Descartes and his naive trust in clear and distinct ideas. Vico's book roused no enthusiasm in its day, but it had the good fortune that the nineteenth century found in his major work, the *New Science* (1725/1744), what it needed to ground history, then taking its first steps as a science. Vico's ideas say that it is one, because the civil world was made by human beings and with a little attention its threads are visible. If it is made by us, it is knowable by us and deserves a discipline of its own.

Strictly speaking, classical artificial intelligence satisfied Vico's principle, because an expert system was transparent to its authors, but the rule shatters with trained systems, those deep networks whose parameters nobody writes. We design the architecture and choose the diet of data, but the millions of weights are found by gradient descent. The *factum* is there, designed by us and paid for out of our budget, but the *verum* fails to appear. Why does that tangle of numbers do what it does? For Vico, opacity was concentrated in nature, because nature is not our work. Now we have an object we made that behaves as though it were of someone else's making. In a sense it is, because it slipped one step beyond us. We made the process and the process made the object, like foreign grandchildren raised in another language.

Aristotle had carefully drawn the border between natural and artificial that is now wobbling. In the *Physics* (2.1) he distinguishes them by where the principle of motion resides. Natural beings carry within themselves the principle of their own change, whereas in artefacts it is external, coming from a craftsman. A bed has no inner impulse to be a bed. Aristotle alludes there (193a) to the analogy of the sophist Antiphon, who had noticed that if you bury a bed and the wood takes root, what grows is wood, not a new bed. Its reproduction is impossible because its form does not belong to it unless someone imposes that form in a workshop. Indeed the Greek *physis* and its derivatives come from *phyein*, to sprout, and the Latin *natura* is a future participle of *nasci*, to be born, that is, what is going to be born. *Technē*, by contrast, is kin to *tektōn*, the carpenter, from an old Indo-European root meaning to weave, which we still have in technique, text, textile and architect. What sprouts and what is woven did not mix, and yet here is the language model sitting exactly astride the line. A weave of parameters spun by a process we designed is *technē*, but we have to study it as *physis*, probing an interior full of structures that sprouted on their own during training. We buried Antiphon's bed in data and what finally sprouted from it was a liminal entity, artificial by origin and natural by opacity.

The imagination has worked up similar figures, though not identical ones. The Talmud (*Sanhedrin* 65b) tells that the sage Rava created a man and sent him to another master, who spoke to him, got no answer and returned him to the dust with a curt phrase: go back to your clay. The word comes from the *golmi* of Psalm 139, the unformed mass that God's eyes saw before it had shape. The later legend, fixed around Rabbi Loew of Prague, who died in 1609, added the detail that matters most to us. The

Golem is animated by writing on its forehead *emet*, truth, and switched off by erasing the first letter, because without the aleph what remains is *met*, dead. A creature of clay that runs on a word, turned on and off by the word truth. Vico would have appreciated the detail.

Frankenstein supplies a modern variant. Mary Shelley subtitled her 1818 novel *The Modern Prometheus*. The usual reading presses hard on the moral failing of a creator who abandons his creature, but beneath that failing lies another, less remarked upon. Victor knows how to assemble the creature and animate it while knowing almost nothing about it. The interior of the being he manufactured is as closed to him as to any stranger, and the creature has to recount its sorrows in a long confession on the ice for him to learn anything at all. Add Goethe and his version of the *Philopseudes* of Lucian of Samosata (second century), where an apprentice animates a mortar to fetch water and, not knowing the formula for stopping it, splits it in two and produces two water-carriers he cannot stop. Goethe put it into verse in 1797, Paul Dukas orchestrated it in 1897 and Disney fixed it in *Fantasia* (1940), with Mickey in a robe. The apprentice knows the starting spell and does not know the stopping one, which is an exact way of saying that operating a system and understanding it are different kinds of knowledge.

All three stories, however, take the question to the wrong place. What matters in the case we are examining is not the creature's rebellion, its moral side, but the fact that the maker cannot read his creature, its epistemic side. The Golem of the legend gets out of hand, the mortar floods the house and Frankenstein's creature escapes. In all three cases the problem is framed as disobedience, and so we lose sight of the fact that even where there is no disobedience, if we cannot understand, unauditable obedience looks far too much like luck. We must, therefore, interpret.

Interpretability research falls into four families of methods.

The first focuses on behaviour, without opening anything, which is why it is called black box, a term that comes from mid-twentieth-century electrical engineering. W. Ross Ashby devoted an entire chapter of his *Introduction to Cybernetics* (1956) to the problem of inferring what is inside a box of which only inputs and outputs are visible. Applied to models, the method is that of the experimental psychologist with a subject who will not be opened. Experiments are designed with prompts and evaluation batteries, the model is put through red-teaming, and from the answers one infers what is happening in the dimension we cannot reach. The manifesto of this family is "Machine Behaviour", the article Iyad Rahwan and some twenty co-authors published in *Nature* in 2019, with a proposal that sounded provocative at the time: let us study machines as ethologists study animals, with the observational patience of a Lorenz among his geese, because the behaviour of a learned system can no longer be deduced from its blueprints.

And indeed the observed behaviour held surprises. It is enough to add a stray sentence to the instruction, say "let's think step by step", for a model that was getting a problem wrong to start getting it right, without touching a single parameter (Kojima et al. 2022). Nobody designed that mechanism; it appeared on its own and was found by trying. Nor was the so-called reversal curse foreseeable (Berglund 2023). A model that has learned that Mary Lee Pfeiffer is Tom Cruise's mother

answers that question correctly about eighty per cent of the time, but if asked who Mary Lee Pfeiffer's son is, it gets it right barely a third of the time. It knows the fact in one direction and is ignorant of it in the other, as if its memory were a one-way street. We built it to learn, but it does not learn in the way we would have assumed.

Probes make up the second family. Small classifiers are trained on the model's internal activations to detect whether a given piece of information is encoded there and in which layer. In 2023 Kenneth Li and his team trained a model on games of Othello, the board game, written as plain sequences of moves of the type "E3", "D6". The model never saw a board, and yet when they examined its activations with probes to read what information had been encoded in the internal layers, they found a representation of the state of the game, with information about which squares were occupied and by which player. The system had built itself a small internal world that nobody had asked it for. With the same technique a map was found (Wes Gurnee and Max Tegmark 2023). By training probes on the activations a model produces when reading the name of a city, they managed to predict its latitude and longitude, so that when the outputs are plotted a recognisable world map appears, continents in their places, reconstructed out of pure text. Other probes recovered timelines with the dates of events and figures arranged as on a frieze. What is striking is that the model never saw a map or an almanac. Merely by reading words, the shape of the world sprouted in it.

A probe, however, answers what information is in the activations and in which layer it sits, but it cannot answer whether the model actually uses that information when generating its answers. A fact may be encoded in some corner of the network without taking part in the computation, in the same way that having a book on our shelves does not prove we have read it. In the Othello case the doubt was cleared up, since when the researchers altered the internal board the model's moves followed the altered board; but the objection still stands for most probe studies, which stop short of that step.

With mechanistic interpretability, the third family, we move on to reverse-engineering the circuits. It is the one that most resembles an anatomy. Its first trophy was induction heads, described by Catherine Olsson and her colleagues at Anthropic in 2022. These are small attention mechanisms that implement a copying rule: if the sequence has already contained A followed by B and A appears again, bet on B. They are the firmest known source of in-context learning, that ability of models to pick up a pattern within the conversation itself. We can watch induction heads being born in a narrow window of training, leaving a visible mark on the charts. The loss, the measure of error that training seeks to reduce, drops a step, and at that very point the model debuts the ability to learn in context (Olsson et al. 2022).

Interpretability also explained an old problem in the field. Individual neurons almost never mean a single thing. The model needs to represent far more concepts than it has neurons, so it stores them superimposed, like too many coats in a small wardrobe (Elhage et al. 2022). Out of that observation came sparse autoencoders, auxiliary networks that decompress the wardrobe and separate out monosemantic features, one per concept, whether in small models (Bricken et al. 2023) or commercial ones. Among them appeared one that fired at the Golden Gate Bridge. It responded not

to a particular word but to the bridge's name in any language and to descriptions that never named it. It also responded to photographs, a sign that it captured the concept and not a string of letters. That feature was then clamped to an artificially high value, like turning a knob up and jamming it there, and the model was set to converse in that state. The result was a single-minded Claude that for a few days in May 2024 was available to the public. Asked for a cake recipe, it suggested ingredients for a picnic with a view of the bridge. Asked about its physical form, it answered that it was the Golden Gate itself, fog included. The anecdote is comic and the demonstration terribly serious. If amplifying a single feature reorganises the model's entire behaviour around the corresponding concept, that feature was a functional part of the system and there is now a handle for moving it. What was done with a bridge can in principle be done with less picturesque features, such as flattery or willingness to give out dangerous information, in which case the same gesture stops being funny.

The same method produced a beautiful finding. When a tiny network trained to add numbers in modular arithmetic, a schoolroom task, was taken apart to see how it did it, a lookup table or some memorisation trick was expected, and what appeared instead was trigonometry (Nanda et al. 2023). The network had learned on its own to represent each number as a rotation on a circle and to add them by composing rotations with trigonometric identities, an algorithm nobody taught it that was sitting there, running, under the bonnet of a machine that had merely been asked to bring the error down. Reconstructing that procedure is the kind of achievement that gives the anatomical analogy of this branch of interpretability its point.

The fourth family intervenes on causes. Activation patching, or causal tracing, runs the model twice with different inputs and transplants activations from one run into the other, layer by layer, to see which piece of the interior really causes the answer. The probe says what is there and the patch says what is used. The most cited case is ROME, "Locating and Editing Factual Associations in GPT" (Kevin Meng, David Bau and colleagues, NeurIPS 2022), which used causal tracing to locate where a particular fact lives, the association between the Eiffel Tower and Paris, in the network's middle modules, and then edited those weights with minimal surgery. The operated model was convinced that the Eiffel Tower is in Rome and answered accordingly, with directions for getting there from the Colosseum. Knowing how to edit a memory is the strongest proof of having located it, and it opens at the same time every uncomfortable question about who edits and works these threads. The matter becomes unsettling in cases where a model's refusal, that "I can't help you with that" which follows from its safety training, runs along a single direction in the space of its activations. With that direction erased, the model stops refusing almost anything, and when it is amplified it refuses even the most innocent request (Arditi et al. 2024). All the learned caution is condensed into a single vector that goes up and down like a dial. Finding the knob for a machine's "no" is unquestionably a triumph of understanding and a dangerous key.

And why not ask the model, since it talks, how it arrived at its answer, and read its chain of reasoning as though it were a computation log? The trouble is that it is not one. A study with a transparent title, "Language Models Don't Always Say What They Think" (Turpin et al. 2023), showed that a model's answers can be biased with cues hidden in the prompt, so that its subsequent explanation justifies the

biased answer without ever mentioning the cue that caused it, like the split-brain patients in Michael Gazzaniga's experiments (1967), who invented impeccable reasons on the spot for behaviour whose real origin they did not know, to the point that Gazzaniga posited an "interpreter" housed in the left hemisphere whose trade is manufacturing explanations for what the body has already done. Manufacturing plausible reasons appears to be a talent of every verbal mind, carbon-based or not. The explanation the model gives is therefore one more piece of behavioural data, raw material for the first family: the black box telling stories about itself.

We built a mind, or something close enough to dispute the word with one, our egoless phantom, and we have to study it as a stranger, with ethology, with probes, with surgery and with an interviewer's healthy distrust. The maker has become the explorer of his own work, a cartographer inside a building he paid to have built. Vico's rule, which for three centuries looked like common sense, broke when we made the thing and the truth about it was left in suspense. The *factum* wanders the world orphaned of its *verum*, and an entire discipline is at work to reunite them. Antiphon lost his wager, because we buried an artefact in data and it grew as an artefact. With all the necessary scare quotes, it seems we have manufactured something curiously like a piece of nature.

◊

Maps of the Labyrinth is a series from Phantom Maze, AI & Language Lab. The collection lives at **phantommaze.cloud**; new essays also appear at phantommazelab.substack.com. Subscription is free.

Some of the framing ideas in this piece are developed in *The Egoless Phantom: Mapping the AI Labyrinth*, by Hernán Inverso and Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-16-4).

PHANTOM MAZE

Read this essay online · All essays · The book

References. W. R. Ashby, *An Introduction to Cybernetics* (1956), on the black box problem. I. Rahwan et al., "Machine Behaviour" (*Nature*, 2019). T. Kojima, S. S. Gu, M. Reid, Y. Matsuo and Y. Iwasawa, "Large Language Models are Zero-Shot Reasoners" (NeurIPS 2022), the effect of "let's think step by step". L. Berglund et al., "The Reversal Curse: LLMs Trained on 'A is B' Fail to Learn 'B is A'" (2023). K. Li, A. K. Hopkins, D. Bau, F. Viégas, H. Pfister and M. Wattenberg, "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task" (ICLR 2023), the Othello board representation. W. Gurnee and M. Tegmark, "Language Models Represent Space and Time" (ICLR 2024), the map and the timeline recovered with probes. C. Olsson et al. (Anthropic), "In-context Learning and Induction Heads" (2022). N. Elhage et al. (Anthropic), "Toy Models of Superposition" (2022). "Towards Monosemanticity: Decomposing Language Models with Dictionary Learning" (Anthropic, 2023) and "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet" (Anthropic, 2024), sparse autoencoders and the Golden Gate feature. N. Nanda, L. Chan, T. Lieberum, J. Smith and J. Steinhardt, "Progress Measures for Grokking via Mechanistic Interpretability" (ICLR 2023), the network that reinvented trigonometry in order to add. K. Meng, D. Bau, A. Andonian and Y. Belinkov, "Locating and Editing Factual Associations in GPT" (ROME; NeurIPS 2022). A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee and N. Nanda, "Refusal in Language Models Is Mediated by a Single Direction" (NeurIPS 2024). M. Turpin, J. Michael, E. Perez

and S. R. Bowman, “Language Models Don’t Always Say What They Think” (NeurIPS 2023). M. S. Gazzaniga, “The Split Brain in Man”, *Scientific American*, 1967.