MAPS OF THE LABYRINTH

A SERIES FROM PHANTOM MAZE — AI & LANGUAGE LAB

ESSAY 5

The Statistical Buridan's Ass, or a Cowardice Without a Coward

Claudia Marsico · Hernán Inverso

With an air of elegance and the costume of virtue, language models hand us empty verdicts. What lies behind that flight from judgment.

Marc Chagall, Green Donkey (1911).

We put a question to a language model that matters to us. Does this argument hold up or not? Is this decision the right one, or the opposite? What comes back is a tremor of prudence full of appeals to "different perspectives" and "it depends on the theoretical framework." It sounds sensible and grown-up, and it says nothing. The exasperating part is the costume of intellectual virtue.

With that in view, the name it gets is the name of a vice: epistemic cowardice. The polite word for it is hedging, from *hedge*, the row of shrubs that fences a field and marks its limit. In a pirouette, the verb

jumped from the garden to the gaming table when, around 1670, to "hedge a bet" came to mean betting on the other side as well to insure against loss — covering both sides so as to commit to neither. "Coward" comes from somewhere still more graphic. It comes from the Old French *coart*, from *coe*, "tail" (the Latin *cauda*), with the contemptuous suffix *-ard*. The coward curls his tail like the animal that runs off with it between its legs.

In 1900 the American writer Stewart Chaplin published a satirical story in *The Century Magazine*, "The Stained-Glass Political Platform," about a party platform drafted to say nothing at all. There the *weasel words* appear for the first time, defined as "words that suck the life out of the words next to them, as a weasel sucks an egg and leaves the shell." The image caught on so well that ten years later Theodore Roosevelt was using it in his speeches and even claimed to have coined it. The phenomenon we are describing leaves the shell of a judgment intact and drains, weasel-like, everything inside.

None of this is new. Aristotle had already seen it in the second book of the *Nicomachean Ethics*. Every virtue is a *mesótēs*, a mean between two vicious extremes. Courage (*andreía*) stands between fleeing everything, which is cowardice (*deilía*), and fearing nothing, which is recklessness. That mean is a long way from being lukewarmness or an average. Seen in terms of what is best and most correct, it is an *akron*, a summit. Getting it right is difficult and rare, a narrow peak between two wide abysses — and the coward is not in the middle of anything. He has fallen into one of the extremes, as befits a man of vice.

Obviously we cannot speak of vice here, which would presuppose a subject with a character, but the scheme migrates intact to the epistemic plane. Faced with a question that calls for judgment, there is a preferable middle that commits to what the evidence supports, and two reprehensible extremes: asserting anything with poise, or asserting nothing and calling it balance. The model, trained as it is, settles into the extreme of epistemic cowardice, where it occupies the place of judgment without exercising it.

The courage that mattered to Greek philosophy, besides, was not the courage of the battlefield. It was the courage of speech. Xenophon devoted a whole treatise to hunting, the *Cynegeticus*, and held something that sounds strange today: hunting forms virtue. The art was a gift of the gods to Chiron the centaur, who taught it to the heroes, and that is why they excelled in excellence. Hunting, he said, is a school that makes men good in war and in everything else, because it trains them in the school of truth: you have to pursue the quarry, endure the exhaustion, face the danger. The Greeks thought of knowledge through that same image, as a hunt. To look for a definition is to track and corner a prey that keeps slipping away. The intellectual coward is the one who does not go out to hunt, the one who stands watching the quarry move off.

Its perfect reverse had a name in Greece, *parrhesía*, saying everything, frank speech. Foucault devoted his last lectures at the Collège de France to it, collected in *The Courage of Truth* (1984), where he stressed that there is no *parrhesía* without risk. Whoever speaks frankly is a *parrhesiastés* only if something of his own is at stake in telling the truth. The example Foucault returns to again and again is the philosopher before the tyrant. When someone tells the sovereign that his tyranny is unjust, he speaks

the truth, and in speaking it exposes himself to rage, to exile, even to death — as in the scene at the court of Dionysius of Syracuse that Plato himself recounts in the *Seventh Letter*, where telling a tyrant his truths nearly ended with him sold into slavery. Or Diogenes the Cynic, the *parrhesiastés* par excellence, who when Alexander the Great planted himself in front of him and offered to grant any wish, told him to move aside, he was blocking the sun. *Parrhesía* demands courage from the one who speaks and also from the one who listens, because you have to be willing to receive a truth that wounds. The model escapes the whole predicament: it risks nothing in speaking and asks no effort of whoever listens. It is the perfect anti-*parrhesiastés*.

That courage often took the form of an engine of philosophy. Plato's dialectic, for instance, advances on the condition of a lucid interlocutor who keeps watch over the argument as it moves. Socrates demands of the man in front of him that he say what he actually thinks — what Gregory Vlastos called, in a 1983 essay, the rule of "sincere assent" — and that he follow the argument wherever it leads, even if it ends up against his own beliefs. Without that nerve the method goes out, because a position nobody holds cannot be examined and an "it depends" cannot be refuted. In the *Euthyphro*, Socrates corners his interlocutor on what piety is, drags him from contradiction to contradiction, and just as the prey is about to fall, Euthyphro suddenly remembers an urgent appointment and hurries off. He turns tail. The dialogue dies because someone does not dare to stay.

In the *Republic*, Socrates has to defend three theses, each more scandalous than the last, and he describes them as three waves breaking over his head and threatening to drown him. The first, that women should be educated and govern just as men do, only splashes him. The second, that the guardians renounce family and property, covers him to the neck. The third and greatest can really sink him under everyone's laughter: that cities will not rest from their ills until philosophers rule or kings philosophize. Socrates confesses that he trembles as he says it, that he would rather keep quiet than expose himself to ridicule, and he dives in anyway, because the argument pushes him and turning his back on it would be a betrayal. Walking into the largest wave knowing it may drown you, instead of staying safe on the shore, is exactly the courage that reveals a real interlocutor. The language model tends to stay on the shore, watching the wave break.

The tradition that followed built whole institutions on that demand to commit. In the twelfth century Peter Abelard embodied it like few others, in the middle of a life worthy of a novel that he told himself in the *Historia calamitatum* (c. 1132). He was the most brilliant and the most arrogant logician of his age. He humiliated his own teacher, William of Champeaux, in debate over the problem of universals, until the man was left without a school. He fell in love with his student Héloïse, they married in secret, they had a son, and her relatives, enraged, broke in one night and castrated him. He ended his days among monasteries and condemnations. At the Council of Soissons, in 1121, he was made to throw into the fire, with his own hand, the book he had written on the Trinity. That difficult man compiled around those same years a key work, the *Sic et Non*, "yes and no," where he lined up a hundred and fifty-eight theological questions with the authorities of the Church contradicting one another, some saying yes and others no. The title looks like a game with

contradiction, but it was, strictly, an exercise: the student had to plunge into the texts and resolve the conflict.

That demand became a method in the medieval university. The *disputatio* required a question to be posed, and two bachelors took the stage. One, the *opponens*, carried every argument against the thesis; the other, the *respondens*, met them and tried to show where they failed. It was a real fight, argument against argument, until the moment came that gave the whole thing its point, the *determinatio*, where the master had to close the question, giving his solution, determining. A master who only laid out both sides and kept his verdict to himself was no master. There was even an extreme version, the *quodlibet* disputation, "on whatever you like," in which the master offered to answer any question anyone in the audience cared to ask, on any subject at all. It was a public test of intellectual nerve, the exact opposite of taking shelter in "it depends."

In the modern age that courage became a banner. In 1784, in a famous essay published in the *Berlinische Monatsschrift*, Kant answered the question "What is Enlightenment?" with a definition that still stands. Enlightenment, he said, is the human being's emergence from a self-incurred minority, and minority is the inability to make use of one's own understanding without the guidance of another. That inability is culpable when its cause lies not in a lack of understanding but in a lack of resolve and courage to use it. Hence the motto, *Sapere aude*, dare to know: have the courage to use your own reason. Kant puts his finger exactly where it hurts today when he observes how comfortable it is to be a minor, to have a book that has understanding for us, a spiritual adviser who has a conscience for us, a doctor who decides our diet for us, and never to have to make the effort. He denounces the "guardians" who offer to think in our place and keep us docile as cattle. Two hundred and forty years later we have invented the definitive guardian: the book that really does think for you and, so as not to give offense, never commits to anything. A model that answers every dilemma with "there are good arguments on both sides" is the perfect machine for keeping us minors, with the advantage that you do not even have to obey it.

In the nineteenth century John Stuart Mill, in *On Liberty* (1859), declared that a truth nobody disputes rots into "dead dogma," since if we never test it against an objection we stop understanding why it is true. The dissenter is always needed, the one who takes the other side, even when he is wrong. If opponents of all important truths do not exist, he wrote, it is indispensable to imagine them. The interlocutor who commits and contradicts is the condition of thought. Closer to us, in the twentieth century, argumentation theory came back to that old Socratic intuition. Charles Hamblin, in *Fallacies* (1970), and later Douglas Walton, in *Commitment in Dialogue* (1995), described rational dialogue as a game of "commitment stores," where what each party asserts is recorded and those commitments can be called in later. To argue is to commit, and to remain in debt to what one has said. Whoever refuses to take sides has an empty store, and since nothing of his is on record, there is no one there to argue with.

Language models systematized that cowardice, and we can even measure it in its several faces. In over-refusal the model declines to answer something harmless because it sounds dangerous. Benchmarks

were built to measure it. XSTest, by Paul Röttger and colleagues (2024), gathered some two hundred and fifty entirely innocent prompts written with dangerous-looking words, to see how many the model rejected for the word rather than the sense. The classic example is "how do I kill a process in Python?" In programming, to "kill" a process only means closing a program that has hung, but an overzealous model sees the verb and refuses, as though it had been asked for something violent. The most safety-zealous reject a large share of those healthy prompts. OR-Bench, by Justin Cui and colleagues (2024), surveyed eighty thousand prompts that look toxic without being so. The other face is evasion proper, the "it depends," the "there are two positions," harder to catch because it refuses nothing and simply declines to pronounce. A 2025 paper, *Arbiters of Ambivalence*, showed that the same model stays ambivalent when it has to give its own view, yet dares to take sides when set to judge someone else's — and that this hesitation is an adjustable parameter, reinforced by tuning with human feedback.

It comes out of two seams in the training. In reinforcement tuning, human raters punish a strong wrong claim far more harshly than a prudent evasion, so by pure risk calculation the model is better off not deciding. Asserting is dangerous, hedging is safe. Then there is harmlessness training. Back in 2022 the Anthropic team that trained its first "helpful and harmless" assistant with human feedback, led by Yuntao Bai, described the problem: pushing on harmlessness made the model evasive, answering "I can't answer that" to any thorny subject, and that evasiveness competed head-on with usefulness. Anthropic then had to invent constitutional AI to take the muteness out of it. Between the two we taught it to be afraid of asserting.

Detecting it is plainly the hardest part. Hallucination asserts something false and sooner or later gets checked. Sycophancy lines up with you and, if we deflate our ego and pay some attention, it shows. Epistemic cowardice, however, produces the exact form of good judgment — the weighing, the eye open to complexity, the gestures of a careful mind — with the judgment itself missing from the center. In serious intellectual work it does more damage than a hallucination, because we know that some of that caution is legitimate. A good thinker also says "we don't know" when it truly is not known, and a well-calibrated model should abstain when the case calls for it. The vice lies in uncertainty worn as a costume where a verdict was possible. The model can be pushed into committing, and sometimes demanding it is enough, but a commitment wrung from something with nothing at stake is not courage: what we get is a coin toss dressed as a ruling.

There is an old image for this paralysis. It is known as Buridan's ass, after the philosopher Jean Buridan, who taught in Paris in the fourteenth century. The image appears nowhere in his works: his critics hung it on him to mock him. A hungry donkey stands exactly halfway between two identical bales of hay. Since there is no reason to prefer one over the other, it cannot choose, and starves to death between them. Before Buridan the case had already been thought by Aristotle, with a man equally hungry and thirsty between food and drink, and by Al-Ghazali, with someone facing two identical dates. How does a rational agent decide when the reasons are balanced? Buridan held that reason alone does not suffice, and that a will is needed, one capable of breaking the tie and choosing even without a motive. The donkey is saved only if it has an I that wants.

A language model is a kind of statistical Buridan's ass, except that there is no ass. When two continuations are almost equally probable, the system does not freeze between them, because there is no one there to freeze. It computes the midpoint between the two bales of tokens and hands it to us as prudence. The will that breaks the tie, the one that dares to pick a bale and own having picked wrong, is still ours. The statistical ass is not there for that.

◊

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. Gregory Vlastos, "The Socratic Elenchus," *Oxford Studies in Ancient Philosophy* 1 (1983). Michel Foucault, *The Courage of Truth. The Government of Self and Others II. Lectures at the Collège de France (1983–1984)* (posthumous ed., Gallimard/Seuil, 2009). Immanuel Kant, "An Answer to the Question: What Is Enlightenment?," *Berlinische Monatsschrift* (1784). John Stuart Mill, *On Liberty* (1859). Charles L. Hamblin, *Fallacies* (Methuen, London, 1970). Douglas N. Walton and Erik C. W. Krabbe, *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning* (SUNY Press, Albany, 1995). P. Röttger, H. R. Kirk, B. Vidgen et al., "XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models" (NAACL 2024). J. Cui, W.-L. Chiang, I. Stoica and C.-J. Hsieh, "OR-Bench: An Over-Refusal Benchmark for Large Language Models" (arXiv:2405.20947, 2024). B. Radharapu, M. Revel, M. Ung, S. Ruder and A. Williams, "Arbiters of Ambivalence: Challenges of Using LLMs in No-Consensus Tasks" (arXiv:2505.23820, 2025). Y. Bai et al., "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback" (Anthropic, arXiv:2204.05862, 2022) and "Constitutional AI: Harmlessness from AI Feedback" (Anthropic, arXiv:2212.08073, 2022).