MAPS OF THE LABYRINTH

A SERIES FROM PHANTOM MAZE — AI & LANGUAGE LAB

ESSAY 3

Hallucination and AI: Phantoms, Psychedelia, and Epistemic Responsibility

Claudia Marsico · Hernán Inverso

The most famous and most feared of AI's failures has gone by some curious names. Follow them and you get the profile of an everyday danger.

Hieronymus Bosch, The Garden of Earthly Delights (c. 1490–1510). A delirium of impossible creatures born of no perception.

In 2023 the Cambridge Dictionary made "hallucinate" its word of the year, in the sense of an AI that produces false information wearing the face of truth. Small wonder. Within months the word had settled into ordinary conversation, and although every new frontier model promises the beast has been tamed, it is still crouched where it always was, waiting for the unwary.

The history of the term is itself hallucinatory. The Latin *alucinari* meant to ramble and lose the thread, to wander off. It was probably related to the Greek *alýein*, to be beside oneself, absent, and to

aláomai, to wander. Its ending, *-ari*, may owe something to *vaticinari*, "to prophesy" and also "to rave," a slippage that joins straying with the raving of someone who talks too much. For centuries, to hallucinate was a wandering of judgment rather than a deception of the senses, and it lived coiled around *phantasía* and *phantasma*, as one sees in Asclepiades of Bithynia. The egoless phantom of our "What Is It That Speaks?" was already entangled with an imagination that could turn hallucinatory.

From there the word adapted itself, with remarkable plasticity, to the most varied senses. It entered English in 1572 through a treatise on ghosts: the translation of the Swiss pastor Ludwig Lavater's book on spirits that walk by night, strange noises, and portents of deaths and the fall of kingdoms. Over the following centuries, that neighborhood of implausible apparitions dragged it toward becoming a name for madness, as it appears in the *Praxis medica* of the anatomist Felix Platter, at the start of the seventeenth century, in his great classification of psychiatric illness.

A little over a hundred years later, in 1712, Joseph Addison, one of the fathers of the English essay and a pioneer of opinion journalism, left behind a curious use that sounds almost like prophecy. In an issue of *The Spectator*, the paper he had founded with Richard Steele, he dismisses a doubtful passage in an ancient text by putting it down to "the hallucination of the transcriber." To hallucinate was the slip of the man copying the tradition by hand, meaning to be faithful, who without noticing introduces something that was not there. Three centuries on, we use the same word for the slip of a machine that recombines that same tradition and smuggles in something foreign.

That line was interrupted. A few decades later the nosologist François Boissier de Sauvages turned the word into a catch-all roomy enough for ringing ears, vertigo, and even sleepwalking. The association with disease is unmistakable in Erasmus Darwin: physician, poet, inventor, grandfather of Charles, and a member of the Lunar Society of Birmingham. Between 1794 and 1796 he published his *Zoonomia; or, The Laws of Organic Life*, an enormous treatise that included a classification of ailments in which *hallucinatio* figures as a deception of the senses, with species such as *hallucinatio visûs*, the eye's, and a *hallucinatio maniacalis* for madness. In those same pages Erasmus sketched ideas of common descent and the transformation of species, decades before his grandson turned them into a theory. Percy Shelley invokes them in the preface he wrote for Mary's *Frankenstein*, to lend plausibility to the reanimation of matter. The catalogue of hallucinations, oddly enough, shares its covers with the germ of evolution and with the monster's scientific alibi.

A few decades later the term seemed to discipline itself, when the French physician Jean-Étienne Esquirol, in his *Des maladies mentales* (1838), gave the name hallucination to the inward conviction of perceiving something with no object to provoke it, and reserved *illusion* for the deformed perception of something that is in fact there. "Visions" struck him as too narrow, since to speak of visions of hearing or of taste is a contradiction in terms. Hallucination, by contrast, made a good generic word. Note that for Esquirol hallucination is a phenomenon of the brain, where perception occurs without a correlate. Where the ancient tradition had put a wandering mind, perception now installs itself. It is a turn that will collide head-on when language models, which perceive nothing whatever, take us back to wandering, this time the wandering of something that does not even have a mind.

Before that, though, look at the twentieth century, where the word found its stickiest association, still with us, with drugs. In 1943 Albert Hofmann took the LSD he had synthesized years earlier, and in 1954 Aldous Huxley recounted his afternoon with mescaline in *The Doors of Perception*. That same year the psychiatrist Abram Hoffer and his colleagues coined "hallucinogen."

The available names painted the experience as illness. "Hallucinogen" tied it to delirium, and "psychotomimetic," the other term in fashion, defined it as an imitation of psychosis. Humphry Osmond, who saw therapeutic promise in these substances, went looking for a word scrubbed clean of that stigma, and asked Huxley for help. Both were men of letters, and they turned the search into a delicious game. They wrote to each other proposing neologisms, each one poured into a rhyming couplet, half joke and half argument. A word that sits well in a line of verse has already won half the battle. Huxley opened with *phanerothyme*, "that which makes the soul manifest":

> *To make this mundane world sublime, Take half a gram of phanerothyme.*

Osmond answered with *psychedelic*, "that which makes the mind manifest":

> *To fathom Hell or soar angelic, Just take a pinch of psychedelic.*

Osmond won, and announced the term in 1957. But "hallucinogen" came through the duel untouched, as the clinical name for those substances, and so "hallucination" stayed welded to the trip and to colors that are not there.

Hieronymus Bosch, The Garden of Earthly Delights *(c. 1490–1510). A delirium of impossible creatures born of no perception.*

With all that baggage on its back, the word turned over and became a compliment. In March 2000, at a conference in Grenoble, Simon Baker and Takeo Kanade, two researchers from Carnegie Mellon, presented a paper with a title that now sounds strange: "Hallucinating Faces." Their algorithm studied thousands of photographs of faces until it had learned what a face looks like up close, and then, from a blurred thumbnail, invented the fine detail that was missing, from the eyes to the grain of the skin, multiplying the pixels by as much as sixty-four. That detail was not in the input image. The model supplied it, out of its knowledge of how faces tend to go. The better the system hallucinated, the sharper and more convincing the photograph came out. The term caught on, and a whole branch of the field adopted it with pride. It is, strictly speaking, the same gesture a language model makes. Only, on a face it is a feat and on a judicial citation it is a scandal. In one case the plausible is enough for us; in the other we need the true.

It did not last, and the compliment turned into an accusation on a fairly precise date. The negative sense, the machine that produces something fluent but unhooked from the real, surfaced around 2017 in machine translation. Philipp Koehn and Rebecca Knowles used it for translations that sounded perfect and had nothing to do with the original. In 2018 Anna Rohrbach and her colleagues fixed it as a technical term while studying systems that describe images: when a model "saw" in a

photograph an object that was not there, they christened it object hallucination. From there it jumped to the rest of text generation, consolidated in the big surveys around 2022, and with the release of ChatGPT that November it passed from the jargon of a few laboratories into everyone's mouth. In barely five years, Rohrbach's technical accusation ended up as the coronation we began with: the word of the year for 2023.

Other names have been put forward. In 2023 three health researchers, Andrew Smith, Felix Greaves, and Trishan Panch, proposed calling it confabulation, a term from neurology that describes the patients of Korsakoff's syndrome. In that condition the capacity to lay down memories is lost, and every hole gets filled with an invented story upheld without a shadow of doubt. With no awareness of the fabrication, patients assemble a memory they do not have. The parallel is suggestive. The machine, too, lacks an acquisition history, an internal time that records when and how firmly it learned each thing, and so everything it says comes out with the same weight, the verified mixed in with the just-improvised. But nobody goes around talking about confabulation. Worse, in the Romance languages, Italian, Spanish, Portuguese, *confabular* pulls toward conspiring, plotting, whispering in corners, which has little to do with what we are trying to name; and elsewhere it sounds relentlessly medical, more so even than the "hallucination" that bothered Osmond and Huxley.

The other name is blunter. In 2024 three philosophers from Glasgow, Michael Hicks, James Humphries, and Joe Slater, titled an article "ChatGPT is bullshit." They had a point. In *On Bullshit* (2005), Frankfurt had distinguished two figures we tend to confuse. The liar knows the truth and works against it, to the point of bending it. The bullshitter could not care less, and only wants what he says to sound good. If it flies, it flies. In that sense the bullshitter is a greater enemy of truth than the liar himself. Hicks and his colleagues add two degrees: hard bullshit, which actively sets out to deceive, and soft bullshit, which asks only for indifference toward the true. Language models would be soft bullshitters, with no relation to truth and with sounding convincing as their single aim. The diagnosis fits better than "hallucination," since it drags in neither mistaken perception nor medicine. But no one is going to stand up at a conference and say the model is bullshitting.

The history is vivid, and yet none of these names, "hallucination" included, explains why the thing is so hard to catch in the act. An invented answer has the same shape as a true one, and gives itself away only from outside, from someone with independent access to the facts. It is the structure of the anamorphosis we discussed in "A Matter of Angles," those figures that assemble only from a certain point of view. What is at stake for whoever does not occupy that point of view, though, is more than a wrong date. In 2023 Steven Schwartz, a New York lawyer with more than thirty years in practice, filed a brief in federal court resting on half a dozen rulings that supported his case. He had asked ChatGPT for them and, with a twinge of doubt, did what seemed prudent: he asked the model itself whether the rulings were real. The model confirmed them one by one. None of them existed. When opposing counsel could not find them, and neither could the judge, Schwartz stood exposed, sanctioned and humiliated after three decades in the profession. He had placed his reputation, and his colleague Peter LoDuca's, entirely in the hands of a text that sounded impeccable, demanding an accounting from the one thing that could not give it.

We love to think this gets fixed with better models, and in part it does, since frontier models hallucinate less than their predecessors. That is exactly where the trap is. The rarer and more plausible the slip becomes, the further we drop our guard, and a terrible one can slide in among thousands of correct and confident answers. An excellent system teaches us to trust it and saves its error for the moment we least expect, like an Edmond Dantès biding his time, except that here nobody is planning anything and we, of course, have done the model no harm. The improvement only hides the danger better.

It can even be shown that the error is unavoidable, with results from learning theory and with Gödel's theorem. No finite system encompasses everything that is true, and at every step a crack remains through which the false gets in, so we can push it down but never to zero. If Esquirol took hallucination out of the eye and spread it across all the senses, language models carried it one step further, into a hallucination with no sense at all, one that runs through words alone. It is the first purely linguistic hallucination, made of the same material with which we tell the truth. The word that began by naming the wandering of a mind found, in the end, a wandering with no mind. To eliminate hallucination would require the model to have a relation to truth. We are nowhere near that.

So what follows? If the model cannot answer for what it asserts, we have to. The responsibility of answering for what one takes to be true cannot be delegated to the model. A Canadian tribunal put that into legal terms in 2024, when it found against Air Canada over a refund policy its chatbot had invented for a passenger. The company tried to distance itself by arguing that the chatbot was a separate entity, and the tribunal would not have it. Nobody gets to hide behind the phantom. Every time we take what it says as good without checking, we adopt its indifference to truth as our own, and we hallucinate along with it. To keep from doing that, we have to hold on to the one part the model will never be able to carry: answering for what we say is true.

◊

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. Douglas Harper, *Online Etymology Dictionary*, 2001, s.v. "hallucinate" and "hallucination"; Jan Dirk Blom, *A Dictionary of Hallucinations*, Springer, 2010; German E. Berrios and Ivana Marková, "The Construction of Hallucination: History and Epistemology," in J. D. Blom and I. Sommer (eds.), *Hallucinations: Research and Practice*, Springer, 2012; Ludwig Lavater, *Of Ghostes and Spirites Walking by Nyght*, London, 1572 (German original, Zurich, 1569); Felix Platter, *Praxis*

medica, Basel, 1602; Joseph Addison, *The Spectator*, no. 470, London, 1712; François Boissier de Sauvages, *Nosologia methodica*, Amsterdam, 1763; Erasmus Darwin, *Zoonomia; or, The Laws of Organic Life*, London, 1794–1796; Jean-Étienne Esquirol, *Des maladies mentales*, Paris, 1838; Aldous Huxley, *The Doors of Perception*, London, 1954; Humphry Osmond, "A Review of the Clinical Effects of Psychotomimetic Agents," *Annals of the New York Academy of Sciences*, New York, 1957; Simon Baker and Takeo Kanade, "Hallucinating Faces," IEEE Intl. Conf. on Automatic Face and Gesture Recognition, Grenoble, 2000; Philipp Koehn and Rebecca Knowles, "Six Challenges for Neural Machine Translation," First Workshop on Neural Machine Translation, 2017; Anna Rohrbach et al., "Object Hallucination in Image Captioning," EMNLP, Brussels, 2018; Ziwei Ji et al., "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys*, 2023; Cambridge Dictionary, "Word of the Year 2023: hallucinate," Cambridge, 2023; Andrew L. Smith, Felix Greaves, and Trishan Panch, "Hallucination or Confabulation? Neuroanatomy as Metaphor in Large Language Models," *PLOS Digital Health*, 2023; Michael Townsen Hicks, James Humphries, and Joe Slater, "ChatGPT is Bullshit," *Ethics and Information Technology*, 2024; Harry Frankfurt, *On Bullshit*, Princeton, Princeton University Press, 2005; Ziwei Xu et al., "Hallucination is Inevitable: An Innate Limitation of Large Language Models," 2024; *Mata v. Avianca*, U.S. District Court (S.D.N.Y.), New York, 2023, and*Moffatt v. Air Canada*, 2024 BCCRT 149, British Columbia, 2024.