

Allucinazione e IA: tra fantasmi, psichedelia e responsabilità epistemica

Claudia Marsico · Hernán Inverso

*Il più famoso e temuto degli errori dell'IA ha avuto nomi curiosi.
Seguirli restituisce il profilo di un pericolo quotidiano.*



Hieronymus Bosch, Il giardino delle delizie (c. 1490–1510). Un delirio di creature impossibili nate da nessuna percezione.

Nel 2023 il dizionario di Cambridge scelse “allucinare” come parola dell’anno, nell’accezione di un’IA che produce informazione falsa con la faccia del vero. C’era poco da stupirsi. In pochi mesi la parola si era installata nella conversazione comune e, benché ogni nuovo modello di frontiera prometta di aver domato la bestia, resta acquattata dov’è sempre stata, in agguato degli incauti.

La storia del termine è essa stessa allucinante. Il latino *alucinari* significava divagare e perdere il filo, come chi se ne va per i rami. Era probabilmente imparentato con il greco *alyein*, essere fuori di sé, assente, e con *aláomai*, vagare. Nella terminazione *-ari* influi forse *vaticinari*, “profetizzare” e anche

“delirare”, in uno slittamento che unisce il vagare al vaneggiare di chi parla troppo. Per secoli allucinare fu uno smarrimento del giudizio più che un inganno dei sensi, e viveva avvolto intorno a *phantasia* e *phantasma*, come si vede in Asclepiade di Bitinia. Il fantasma senza ego del nostro “Che cos’è quello che parla” era già intrecciato con un’immaginazione che poteva farsi allucinatoria.

Da lì la parola si adattò con plasticità ai sensi più vari. In inglese entrò nel 1572 attraverso un trattato sui fantasmi, la traduzione del libro del pastore svizzero Ludwig Lavater sugli spiriti che camminano di notte, i rumori strani e i presagi di morti e di caduta di regni. Con i secoli, quel vicinato di apparizioni inverosimili la trascinò a diventare un nome della follia, come appare nella *Praxis medica* dell’anatomista Felix Platter, all’inizio del Seicento, nella sua grande classificazione delle malattie psichiatriche.

Poco più di un secolo dopo, nel 1712, Joseph Addison, uno dei padri del saggio inglese e pioniere del giornalismo d’opinione, lasciò un uso curioso che suona quasi come una profezia. In un numero di *The Spectator*, il foglio che aveva fondato con Richard Steele, liquida un passo dubbio di un testo antico attribuendolo a “un’allucinazione del copista”. Allucinare era lo scivolone di chi copia a mano la tradizione volendo esserle fedele e, senza accorgersene, introduce qualcosa che non c’era. Tre secoli dopo chiamiamo allo stesso modo lo scivolone di una macchina che ricombina quella stessa tradizione e vi infila qualcosa di estraneo.

Quella linea si interruppe. Qualche decennio più tardi il nosologo François Boissier de Sauvages trasformò la parola in un grande contenitore dove entravano il ronzio d’orecchi, la vertigine e perfino il sonnambulismo. L’associazione con la malattia è chiarissima in Erasmus Darwin: medico, poeta, inventore, nonno di Charles e membro della Lunar Society di Birmingham. Fra il 1794 e il 1796 pubblicò la sua *Zoonomia, ovvero le leggi della vita organica*, un trattato enorme che comprendeva una classificazione dei mali in cui *hallucinatio* figura come inganno dei sensi, con specie quali la *hallucinatio visus*, quella dell’occhio, e una *hallucinatio maniacalis* per la follia. In quelle stesse pagine Erasmus abbozzò idee di discendenza comune e di trasformazione delle specie, decenni prima che suo nipote ne facesse una teoria. Percy Shelley le richiama nella prefazione che scrisse per il *Frankenstein* di Mary, per dare verosimiglianza alla rianimazione della materia. Curiosamente, il catalogo delle allucinazioni divide la copertina con il germe dell’evoluzione e con l’alibi scientifico del mostro.

Qualche decennio dopo il termine parve disciplinarsi, quando il medico francese Jean-Étienne Esquirol, in *Des maladies mentales* (1838), chiamò allucinazione la convinzione intima di percepire qualcosa senza che alcun oggetto la provochi, e riservò illusione alla percezione deformata di qualcosa che c’è davvero. “Visioni” gli sembrava troppo stretto, perché parlare di visioni dell’udito o del gusto è un controsenso. Allucinazione, invece, era un buon termine generico. Si noti che per Esquirol l’allucinazione è un fenomeno del cervello, dove si dà percezione senza correlato. Dove la tradizione antica metteva la mente vagabonda si installa la percezione: una svolta destinata a scontrarsi frontalmente quando i modelli linguistici, che non percepiscono nulla, ci riporteranno al vagabondare, stavolta di qualcosa che non ha nemmeno una mente.

Guardiamo però prima al Novecento, dove la parola trovò la sua associazione più tenace, tuttora viva, con le droghe. Nel 1943 Albert Hofmann provò l'LSD che aveva sintetizzato anni prima, e nel 1954 Aldous Huxley raccontò il suo pomeriggio con la mescalina in *Le porte della percezione*. Quello stesso anno lo psichiatra Abram Hoffer e i suoi colleghi coniarono “allucinogeno”.

I nomi disponibili dipingevano l'esperienza come malattia. “Allucinogeno” la legava al delirio, e “psicotomimetico”, l'altro termine in voga, la definiva come un'imitazione della psicosi. Humphry Osmond, che in quelle sostanze vedeva un potenziale terapeutico, si mise a cercare una parola ripulita da quello stigma e chiese aiuto a Huxley. Erano entrambi uomini di lettere e trasformarono la ricerca in un gioco delizioso. Si scrissero proponendosi neologismi, ciascuno versato in un distico, mezzo per scherzo e mezzo sul serio. Una parola che sta bene in un verso ha già vinto metà della battaglia. Huxley aprì con *phanerothyme*, “ciò che manifesta l'anima”: *To make this mundane world sublime, / Take half a gram of phanerothyme* (“per rendere sublime questo mondo mondano, prendi mezzo grammo di fanerotima”). Osmond rispose con *psychedelic*, “ciò che manifesta la mente”: *To fathom Hell or soar angelic, / Just take a pinch of psychedelic* (“per sondare l'Inferno o librarti angelico, prendi un pizzico di psichedelico”). Vinse Osmond, che annunciò il termine nel 1957, ma “allucinogeno” attraversò il duello indenne come nome clinico di quelle sostanze, e così “allucinazione” restò saldata al viaggio e ai colori che non ci sono.

Con tutto quel bagaglio sulle spalle, la parola si capovolse e diventò un elogio. Nel marzo del 2000, a un convegno di Grenoble, Simon Baker e Takeo Kanade, due ricercatori della Carnegie Mellon, presentarono un lavoro dal titolo che oggi suona strano, *Hallucinating Faces*, “allucinare volti”. Il loro algoritmo studiava migliaia di fotografie di volti finché non imparava com'è fatto un volto da vicino, e poi, a partire da una miniatura sfocata, inventava il dettaglio fine che mancava, dagli occhi alla grana della pelle, moltiplicando i pixel fino a sessantaquattro volte. Quel dettaglio non era nell'immagine di partenza: lo forniva il modello, con il suo sapere su come sono fatti di solito i volti. Meglio il sistema allucinava, più nitida e convincente veniva la foto. Il termine attecchì e un intero ramo della disciplina lo adottò con orgoglio. È, a rigore, lo stesso gesto che compie un modello linguistico. Solo che su un volto è un'impresa e su una citazione giudiziaria è uno scandalo. In un caso ci basta il plausibile; nell'altro ci serve il vero.

Non durò, e l'elogio si fece accusa in una data abbastanza precisa. Il senso negativo, la macchina che produce qualcosa di fluido ma sganciato dal reale, affiorò verso il 2017 nella traduzione automatica. Philipp Koehn e Rebecca Knowles lo usarono per traduzioni che suonavano perfette e non avevano nulla a che vedere con l'originale. Nel 2018 Anna Rohrbach e i suoi colleghi lo fissarono come termine tecnico studiando i sistemi che descrivono immagini: quando un modello “vedeva” in una foto un oggetto che non c'era, lo battezzarono *object hallucination*. Di lì saltò al resto della generazione di testo, si consolidò nelle grandi rassegne verso il 2022 e, con l'uscita di ChatGPT quel novembre, passò dal gergo di pochi laboratori alla bocca di tutti. In appena un lustro l'accusa tecnica di Rohrbach finì convertita nell'incoronazione da cui siamo partiti: la parola dell'anno 2023.

Sono stati proposti altri nomi. Nel 2023 tre ricercatori in ambito sanitario, Andrew Smith, Felix Greaves e Trishan Panch, proposero di chiamarla confabulazione, un termine della neurologia che descrive i pazienti della sindrome di Korsakoff. In quel quadro si perde la capacità di fissare i ricordi e ogni buco viene riempito con un racconto inventato, sostenuto senza ombra di dubbio. Senza coscienza della fabulazione, i pazienti costruiscono una memoria che non hanno. Il parallelo è suggestivo. Anche alla macchina manca una storia di acquisizione, un tempo interno che registri quando e con quanta fermezza ha imparato ogni cosa, e così tutto quello che dice esce con lo stesso peso, il verificato mescolato all'appena improvvisato. Ma nessuno va in giro a parlare di confabulazione. Peggio ancora, nelle lingue romanze, italiano, spagnolo, portoghese, “confabulare” tira verso il cospirare, il tramare, il bisbigliare negli angoli, che poco ha a che fare con ciò che si vuole nominare; e altrove suona implacabilmente medico, più ancora dell'*hallucination* che dava fastidio a Osmond e Huxley.

L'altro nome è più brusco. Nel 2024 tre filosofi di Glasgow, Michael Hicks, James Humphries e Joe Slater, intitolarono un articolo “ChatGPT is bullshit”. Un punto ce l'avevano. In *Stronzate* (2005), Frankfurt aveva distinto due figure che tendiamo a confondere. Il bugiardo conosce la verità e lavora contro di essa, fino a torcerla. A chi dice stronzate non importa nulla, e vuole soltanto che ciò che dice suoni bene. Se passa, passa. In questo senso è un nemico della verità più grande del bugiardo stesso. Hicks e i suoi colleghi aggiungono due gradi: la stronzata dura, che punta attivamente a ingannare, e quella molle, che chiede solo indifferenza verso il vero. I modelli linguistici sarebbero produttori molli di stronzate, senza alcun rapporto con la verità e con il suonare convincenti come unico scopo. La diagnosi calza meglio di “allucinazione”, perché non si porta dietro né percezioni sbagliate né medicina. Ma nessuno si alzerà a un convegno per dire che il modello sta dicendo stronzate.

La storia è vivace, e tuttavia nessuno di questi nomi, “allucinazione” compresa, spiega perché la cosa sia tanto difficile da cogliere sul fatto. Una risposta inventata ha la stessa forma di una vera e si tradisce solo dall'esterno, da parte di chi ha accesso indipendente ai fatti. È la struttura dell'anamorfosi di cui abbiamo parlato in “Questione di angoli”, quelle figure che si compongono solo da un certo punto di vista. Ciò che è in gioco per chi non occupa quel punto di vista, però, è più di una data sbagliata. Nel 2023 Steven Schwartz, un avvocato di New York con più di trent'anni di professione, depositò davanti a un tribunale federale una memoria fondata su una mezza dozzina di sentenze che gli davano ragione. Le aveva chieste a ChatGPT e, con una punta di dubbio, fece ciò che sembrava prudente: chiese al modello stesso se quelle sentenze fossero reali. Il modello glielne confermò una per una. Nessuna esisteva. Quando la controparte non riuscì a trovarle, e nemmeno il giudice, Schwartz restò esposto, sanzionato e umiliato dopo tre decenni di mestiere. Aveva messo la propria reputazione, e quella del collega Peter LoDuca, interamente nelle mani di un testo che suonava impeccabile, chiedendo conto all'unica cosa che non poteva dargliene.

Ci piace pensare che tutto si risolva con modelli migliori, e in parte è vero, perché i modelli di frontiera allucinano meno dei loro predecessori. Ed è esattamente lì la trappola. Più lo scivolone diventa raro e plausibile, più abbassiamo la guardia, e uno terribile può infilarsi fra migliaia di risposte corrette e sicure. Un sistema eccellente ci insegna a fidarci e conserva il suo errore per il momento in cui meno ce

lo aspettiamo, come un Edmond Dantès che rimanda la vendetta, salvo che qui nessuno sta pianificando nulla e noi, ovviamente, al modello non abbiamo fatto alcun male. Il miglioramento si limita a nascondere meglio il pericolo.

Si può perfino dimostrare che l'errore è ineludibile, con risultati della teoria dell'apprendimento e con il teorema di Gödel. Nessun sistema finito abbraccia tutto ciò che è vero, e a ogni passo resta una fessura da cui entra il falso, così che possiamo spingerlo verso il basso ma mai fino a zero. Se Esquirol tolse l'allucinazione dall'occhio per distribuirla fra tutti i sensi, i modelli linguistici l'hanno portata un passo più in là, verso un'allucinazione senza alcun senso, che passa soltanto per le parole. È la prima allucinazione puramente linguistica, fatta dello stesso materiale con cui diciamo la verità. La parola che aveva cominciato nominando il divagare di una mente ha trovato, alla fine, un divagare senza mente. Eliminare l'allucinazione richiederebbe che il modello avesse un rapporto con la verità. Non ci siamo neanche vicini.

E allora? Se il modello non può rispondere di ciò che afferma, dobbiamo farlo noi. La responsabilità di rispondere di ciò che si dà per vero non si delega al modello. Un tribunale canadese lo ha tradotto in termini giuridici nel 2024, condannando Air Canada per una politica di rimborso che il suo chatbot aveva inventato a un passeggero. L'azienda tentò di smarcarsi sostenendo che il chatbot fosse un'entità a parte, e il tribunale non l'ha accettato. Nessuno può nascondersi dietro il fantasma. Ogni volta che diamo per buono ciò che dice senza verificarlo, adottiamo come nostra la sua indifferenza verso la verità e alluciniamo con lui. Per non farlo dobbiamo tenerci la parte che il modello non potrà mai portare: rispondere di ciò che diciamo essere vero.



Mappe del labirinto è una serie di Phantom Maze, AI & Language Lab. La collezione vive su phantommaze.cloud. L'iscrizione è gratuita.

Alcune delle idee che inquadrano questo testo sono sviluppate in *Il fantasma senza ego. Una mappa nel labirinto dell'intelligenza artificiale*, di Hernán Inverso e Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-19-5).

PHANTOM MAZE

[Leggi questo saggio online](#) · [Tutti i saggi](#) · [Il libro](#)

Riferimenti. Douglas Harper, *Online Etymology Dictionary*, 2001, s.v. “hallucinate” e “hallucination”; Jan Dirk Blom, *A Dictionary of Hallucinations*, Springer, 2010; German E. Berrios e Ivana Marková, “The Construction of Hallucination: History and Epistemology”, in J. D. Blom e I. Sommer (a cura di), *Hallucinations: Research and Practice*, Springer, 2012; Ludwig Lavater, *Of Ghostes and Spirites Walking by Nyght*, Londra, 1572 (orig. tedesco, Zurigo, 1569); Felix Platter, *Praxis medica*, Basilea, 1602; Joseph Addison, *The Spectator*, n. 470, Londra, 1712; François Boissier de Sauvages, *Nosologia methodica*, Amsterdam, 1763; Erasmus Darwin, *Zoonomia; or, The Laws of Organic Life*, Londra, 1794–1796; Jean-Étienne Esquirol, *Des mala-*

dies mentales, Parigi, 1838; Aldous Huxley, *The Doors of Perception*, Londra, 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 e Takeo Kanade, “Hallucinating Faces”, IEEE Intl. Conf. on Automatic Face and Gesture Recognition, Grenoble, 2000; Philipp Koehn e 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, Bruxelles, 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 e Trishan Panch, “Hallucination or Confabulation? Neuroanatomy as Metaphor in Large Language Models”, *PLOS Digital Health*, 2023; Michael Townsen Hicks, James Humphries e Joe Slater, “ChatGPT is Bullshit”, *Ethics and Information Technology*, 2024; Harry Frankfurt, *On Bullshit*, 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, e *Moffatt v. Air Canada*, 2024 BCCRT 149, British Columbia, 2024.