

Alucinación e IA: entre fantasmas, psicodelia y responsabilidad epistémica

Claudia Marsico · Hernán Inverso

El más famoso y temido de los errores de IA tuvo nombres curiosos que nos sirven para trazar el perfil de ese peligro cotidiano.



Hieronymus Bosch, El jardín de las delicias (c. 1490–1510). Un delirio de criaturas imposibles nacidas de ninguna percepción.

En 2023 el diccionario de Cambridge eligió “alucinar” como palabra del año, en su acepción de una IA que produce información falsa con cara de verdadera. No es para menos. En unos meses la palabra se había instalado en la conversación común y aunque cada nuevo modelo de frontera prometa que controló a la bestia, sigue tan agazapada como siempre a la caza de incautos.

La historia del término es en sí misma alucinante. El latín *alucinari* refería a divagar y perderse, como quien se va por las ramas. Probablemente estaba emparentado con el griego *alýein*, andar fuera de sí, como ausente, y con *aláomai*, vagar. En su terminación *-ari* acaso influyó *vaticinari*, “profetizar” y

también “delirar”, en un descontrol que junta el vagar y el desvarío del que habla de más. Durante siglos, alucinar fue un extravío del juicio antes que un engaño de los sentidos, que estaba asociado, como se ve en Asclepiades de Bitinia, a *phantasia* y *phantasma*. El fantasma sin ego de nuestro “Qué es eso que habla” ya vivía enroscado con la imaginación que podía ser alucinatoria.

Desde ahí la palabra se adaptó plásticamente a sentidos variadísimos. Al inglés entró en 1572 a través de un tratado de fantasmas, la traducción del libro del pastor suizo Ludwig Lavater sobre espíritus que caminan de noche y ruidos extraños, con presagios de muertes y caídas de reinos. Con los siglos, el entorno de las apariciones inverosímiles la arrastró a convertirse en nombre de la locura, como aparece en el *Praxis medica* del anatomista Felix Plater a principios del s. XVII en su gran clasificación de las enfermedades psiquiátricas.

Pasado algo más de un siglo, en 1712, Joseph Addison, uno de los padres del ensayo inglés y pionero del periodismo de opinión, legó un curioso uso que casi suena a profecía. En un número de *The Spectator*, que había fundado con Richard Steele, descarta un pasaje dudoso de un texto antiguo diciendo que se debía seguramente a una “alucinación del copista”. Alucinar era el desliz de quien copia a mano la tradición pretendiendo ser fiel y, sin advertirlo, introduce algo que no estaba. Tres siglos después llamamos igual al desliz de una máquina que recombina esa misma tradición y cuela algo ajeno.

Esta vía se interrumpió y unas décadas después el nosólogo François Boissier de Sauvages la convirtió en un gran cajón de sastre donde cabían desde el zumbido de oídos y el vértigo hasta el sonambulismo. Esa asociación con las enfermedades es clarísima en Erasmus Darwin, médico, poeta e inventor, abuelo de Charles y miembro de la Lunar Society de Birmingham. Entre 1794 y 1796 publicó su *Zoonomia o Las leyes de la vida orgánica*, un tratado enorme que incluía una clasificación de dolencias donde *hallucinatio* aparece como engaño de los sentidos, con especies como la *hallucinatio visus*, la del ojo y una *hallucinatio maniacalis* para la locura. En esas mismas páginas Erasmus esbozó ideas de descendencia común y de transformación de las especies, décadas antes de que su nieto las volviera teoría. A ellas remite Percy Shelley en el prólogo que hizo para el *Frankenstein* de Mary para dar verosimilitud a la reanimación de la materia. Curiosamente, el catálogo de alucinaciones comparte tapas con el germen de la evolución y con la coartada científica del monstruo.

Unas décadas después, el término pareció disciplinarse cuando el médico francés Jean-Étienne Esquirol, en su tratado *Des maladies mentales* (1838), llamó alucinación a la convicción íntima de percibir algo sin que ningún objeto lo provoque y dejó ilusión para la percepción deformada de algo que sí está. El término “visiones” le resultaba estrecho, porque hablar de visiones del oído o el gusto es un contrasentido. Alucinación era, en cambio, un buen término genérico. Notemos que para Esquirol la alucinación es un fenómeno del cerebro, donde se da la percepción sin correlato. Donde la tradición antigua ponía la mente vagabunda se instala la percepción, en un viraje que va a chocar de frente cuando los modelos de lenguaje, que no perciben nada, nos lleven de nuevo al vagabundeo, pero esta vez de algo que ni siquiera tiene mente.

Hieronymus Bosch, El jardín de las delicias (c. 1490–1510). Un delirio de criaturas imposibles nacidas de ninguna percepción.

Pero miremos antes el siglo XX, donde la palabra encontró su asociación pegajosa, todavía vigente, con las drogas. En 1943 Albert Hofmann probó el LSD que había sintetizado años antes y en 1954 Aldous Huxley contó su tarde con mescalina en *Las puertas de la percepción*. Ese mismo año, los psiquiatras Abram Hoffer y sus colegas acuñaron “alucinógeno”.

Los nombres disponibles pintaban la experiencia como enfermedad. “Alucinógeno” la ataba al delirio y “psicotomimético”, el otro término de moda, la definía como un remedo de psicosis. Humphry Osmond, que le veía a estas sustancias un potencial terapéutico, comenzó a buscar una palabra limpia de ese estigma y le pidió ayuda a Huxley. Los dos eran hombres de letras y convirtieron la búsqueda en un juego delicioso. Se cartearon proponiéndose neologismos y cada uno vertió el suyo en versos pareados, medio en chiste y medio en serio. En rigor, una palabra que suena bien en una poesía ya tiene media batalla ganada. Huxley abrió con *phanerothyme*, “lo que manifiesta el alma”: *To make this mundane world sublime, / Take half a gram of phanerothyme* (“para volver sublime este mundo mundano, tomá medio gramo de fanerotima”). Osmond le respondió con *psychedelic*, “lo que manifiesta la mente”: *To fathom Hell or soar angelic, / Just take a pinch of psychedelic* (“para sondear el infierno o remontar angélico, toma una pizca de psicodélico”). Ganó Osmond, que anunció el término en 1957, pero “alucinógeno” siguió inmune a este combate como el nombre clínico de esas sustancias, de modo que “alucinación” quedó soldada al viaje y a los colores que no están.

Con todo ese equipaje encima, la palabra dio un vuelco y llegó a ser un elogio. En marzo de 2000, en un congreso en Grenoble, Simon Baker y Takeo Kanade, dos investigadores de Carnegie Mellon, presentaron un trabajo con un título que hoy suena raro, *Hallucinating Faces*, “alucinando caras”. Su algoritmo estudiaba miles de fotos de rostros hasta aprender cómo es una cara de cerca, y después, a partir de una miniatura borrosa, inventaba los detalles finos que le faltaban, desde los ojos hasta el grano de la piel, multiplicando los píxeles hasta por sesenta y cuatro. Esos detalles no estaban en la imagen de entrada y los reponía el modelo, con su saber sobre cómo suelen ser las caras. Cuanto mejor alucinaba el sistema, más nítida y convincente salía la foto. El término prendió y toda una rama de la disciplina lo adoptó con orgullo. En rigor, es el mismo gesto que hace un modelo de lenguaje, solo que sobre una cara es una proeza y sobre una cita judicial resulta un escándalo. En un caso nos alcanza con lo plausible, en el otro necesitamos lo verdadero.

No duró mucho y pronto el elogio se volvió acusación en una fecha bastante precisa. El sentido negativo asociado con la máquina que produce algo fluido pero desconectado de lo real apareció hacia 2017 en la traducción automática. Philipp Koehn y Rebecca Knowles lo usaron para dar cuenta de traducciones que sonaban perfectas y no tenían nada que ver con el original. En 2018, Anna Rohrbach y sus colegas lo fijaron como término técnico estudiando los sistemas que describen imágenes, cuando un modelo “veía” en una foto un objeto que no estaba y lo bautizaron *object hallucination*. De ahí saltó al resto de la generación de texto, se consolidó en los grandes relevamientos hacia 2022 y, con la salida de ChatGPT ese noviembre, pasó de la jerga de unos pocos laboratorios a la boca

de todos. En apenas un lustro, la acusación técnica de Rohrbach terminó convertida en la coronación con la que empezamos: la palabra del año 2023.

Aquí también hay otros nombres candidatos. En 2023, tres investigadores en salud, Andrew Smith, Felix Greaves y Trishan Panch, propusieron llamarla confabulación, un término de la neurología que describe a los pacientes del síndrome de Korsakoff. En este cuadro se pierde la capacidad de fijar recuerdos y se rellena cada hueco con un relato inventado sostenido sin sombra de duda. Sin conciencia de la fabulación, los pacientes arman una memoria que no tienen. El paralelo es sugerente. A la máquina también le falta una historia de adquisición, un tiempo interno que registre cuándo y con qué firmeza aprendió cada cosa y entonces todo lo que dice sale con el mismo peso, mezclando lo verificado con lo improvisado. Pero nadie anda por ahí hablando de confabulación. Para peor, lenguas romances como el italiano, el español y el portugués asocian “confabular” con “conspirar”, “maquinar”, “cuchichear”, que poco tienen que ver con lo que se quiere nombrar, y suena en otras como muy médico, más todavía que el *hallucination* que molestaba a Osmond y Huxley.

El otro nombre es más brusco. En 2024, tres filósofos de Glasgow, Michael Hicks, James Humphries y Joe Slater, titularon un artículo “ChatGPT is bullshit”. Tenían un punto. En su ensayo *On Bullshit* (2005), Frankfurt había distinguido dos figuras que solemos confundir. El mentiroso conoce la verdad y trabaja contra ella hasta el punto de torcerla. Al que dice macanas, en cambio, le da exactamente igual y solo quiere que lo que dice suene bien. Si pasa, pasa. En este sentido, el macaneador es un enemigo mayor de la verdad que el propio mentiroso. Hicks y sus colegas le agregan dos grados: la macana dura, que busca engañar de manera activa, y la macana blanda, que solo pide despreocupación por lo verdadero. Los modelos de lenguaje serían como macaneadores blandos sin relación con la verdad y con lo que suena convincente como único objetivo. Ese diagnóstico le calza mejor que “alucinación”, porque no trae a cuento percepciones equivocadas ni medicina, pero nadie va a andar por ahí diciendo bullshit o macanas.

Esta historia es vistosa, pero lo cierto es que ninguno de esos nombres, ni tampoco “alucinación”, explica por qué esta cosa es tan difícil de atrapar en el acto. Una respuesta inventada tiene la misma forma que una verdadera y se delata solo desde afuera, desde quien tiene acceso propio a los hechos. Es la estructura de la anamorfosis de la que hablamos en *Cuestión de ángulos*, con esas figuras que solo se arman desde cierto punto de vista. Sin embargo, lo que se juega quien no ocupa ese punto de vista es más que un dato equivocado o perderse un “escorzo anamórfico”. En 2023, Steven Schwartz, un abogado de Nueva York con más de treinta años de ejercicio, presentó ante un tribunal federal un escrito apoyado en media docena de fallos que le daban la razón. Se los había pedido a ChatGPT y, con una punzada de duda, hizo lo que parecía prudente: le preguntó al propio modelo si esos fallos eran reales. El modelo se los confirmó uno por uno, pero ninguno existía. Cuando la contraparte no logró encontrarlos y el juez tampoco, Schwartz quedó expuesto ante todos, sancionado y humillado después de tres décadas de oficio. Había puesto su reputación y la de su colega Peter LoDuca enteras en manos de un texto que sonaba impecable, pidiéndole cuentas a algo que no podía dárselas.

Nos encanta pensar que esto se arregla con mejores modelos y en parte es verdad, porque los de frontera alucinan menos que sus antecesores, pero ahí mismo está la trampa. Cuanto más raro y más verosímil se vuelve el desliz, más bajamos la guardia. Se puede colar uno terrible entre miles de respuestas correctas y seguras. Un sistema excelente nos enseña a confiar y guarda su error para el momento en que menos lo esperamos, como un Edmond Dantès que posterga la venganza, solo que aquí no hay nadie planificando y nosotros, por supuesto, no le hicimos ningún daño al modelo. Lo cierto es que la mejora solo esconde mejor el peligro.

Se puede incluso probar que el error es ineludible con resultados de la teoría del aprendizaje y hasta con el teorema de Gödel. Ningún sistema finito abarca todo lo verdadero y en cada paso queda una hendija por donde entra lo falso, de modo que podemos empujarlo hacia abajo, pero nunca hasta cero. Si Esquirol sacó la alucinación del ojo para repartirla entre todos los sentidos, los modelos de lenguaje la llevaron un paso más lejos, hacia una alucinación sin ningún sentido, que pasa solo por las palabras. Es la primera alucinación puramente lingüística, hecha del mismo material con que decimos la verdad. La palabra que empezó nombrando el divagar de una mente encontró, al final, un divagar sin mente. Eliminar la alucinación exigiría que el modelo tuviera una relación con la verdad. No estamos ni cerca de ahí.

¿Y entonces? Si el modelo no puede hacerse cargo de lo que afirma, tenemos que hacerlo nosotros. La responsabilidad de responder por lo que uno da por cierto no se puede delegar en el modelo. Un tribunal canadiense lo tradujo a términos legales en 2024, cuando condenó a Air Canada por una política de reembolso que su chatbot le inventó a un pasajero. La empresa quiso desligarse alegando que el chatbot era una entidad aparte y el tribunal no lo aceptó. Nadie puede escudarse en el fantasma. Cada vez que damos por bueno lo que dice sin verificarlo, adoptamos como propia su indiferencia por la verdad y alucinamos con él. Para no hacerlo tenemos que quedarnos con la parte que el modelo nunca va a poder cargar, que es la de responder por lo que decimos que es cierto.

Mapas del laberinto es una serie de Phantom Maze, AI & Language Lab. La colección vive en phantommaze.cloud; los ensayos aparecen también en phantommazeilabes.substack.com. La suscripción es gratuita.

Algunas de las ideas que enmarcan este texto se desarrollan en *El fantasma sin ego. Un mapa en el laberinto de la inteligencia artificial*, de Hernán Inverso y Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-10-2).

PHANTOM MAZE

[Leer este ensayo en línea](#) · [Todos los ensayos](#) · [El libro](#)

Referencias. Douglas Harper, *Online Etymology Dictionary*, 2001, s.v. "hallucinate" y "hallucination"; Jan Dirk Blom, *A Dictionary of Hallucinations*, Springer, 2010; German E. Berrios e Ivana Marková, "The construction of hallucination: history and epistemology", en J. D. Blom e I. Sommer (eds.), *Hallucinations: Research and Practice*, Springer, 2012; Ludwig Lavater, *Of Ghostes and Spirites Walking by Nyght*, Londres, 1572 (orig. alemán, Zúrich, 1569); Felix Platter, *Praxis medica*, Basilea, 1602; Joseph Addison, *The Spectator*, n.º 470, Londres, 1712; François Boissier de Sauvages, *Nosologia methodica*, Ámster-

dam, 1763; Erasmus Darwin, *Zoonomia; or, The Laws of Organic Life*, Londres, 1794–1796; Jean-Étienne Esquirol, *Des maladies mentales*, París, 1838; Aldous Huxley, *The Doors of Perception*, Londres, 1954; Humphry Osmond, "A Review of the Clinical Effects of Psychotomimetic Agents", *Annals of the New York Academy of Sciences*, Nueva York, 1957; Simon Baker y Takeo Kanade, "Hallucinating Faces", IEEE Intl. Conf. on Automatic Face and Gesture Recognition, Grenoble, 2000; Philipp Koehn y Rebecca Knowles, "Six Challenges for Neural Machine Translation", First Workshop on Neural Machine Translation, 2017; Anna Rohrbach y col., "Object Hallucination in Image Captioning", EMNLP, Bruselas, 2018; Ziwei Ji y col., "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 y Trishan Panch, "Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models", *PLOS Digital Health*, 2023; Michael Townsen Hicks, James Humphries y Joe Slater, "ChatGPT is Bullshit", *Ethics and Information Technology*, 2024; Harry Frankfurt, *On Bullshit*, Princeton, Princeton University Press, 2005; Ziwei Xu y col., "Hallucination is Inevitable: An Innate Limitation of Large Language Models", 2024; *Mata v. Avianca*, U.S. District Court (S.D.N.Y.), Nueva York, 2023 y *Moffatt v. Air Canada*, 2024 BCCRT 149, Columbia Británica, 2024.