Título La Ilusión de la Objetividad en IA Generativa: De los Sesgos y Alucinaciones a la Construcción de Realidades Alternas como consecuencia de su deficiencia de Calidad del Software

Autor

Prof. Gustavo Aldegani, Facultad de Ingeniería y Tecnología Informática, Universidad de Belgrano

Abstract

Las Inteligencias Artificiales Generativas se presentan a menudo como herramientas neutrales y objetivas. Este artículo desafía esa narrativa, argumentando que los sesgos inherentes en sus datos de entrenamiento y su propensión a la "alucinación" no son meros errores técnicos, sino características estructurales que las convierten en constructores de realidades inexactas o equivocadas. En contextos de incertidumbre social, donde la demanda de información veraz es crítica para la toma de buenas decisiones, el efecto de estas fallas se amplifica, aportando distorsiones a la visión de la realidad objetiva. A partir del análisis inicial se exploran sus impactos sociales y las dificultades de atribución de responsabilidad y finalmente se presentan marcos de referencia como ISO/IEC 25010, MyFEPS y OWASP AI Security & Privacy Guide como alternativas de solución para la evaluación y auditoría de la calidad de estos sistemas, que básicamente son desarrollos de software.

Palabras clave en español:

- Inteligencia Artificial Generativa
- Impacto social de la IA
- Alucinaciones en IA
- Sesgos algorítmicos
- Auditoría de calidad de software

Título en (English) The Illusion of Objectivity in Generative AI: From Biases and Hallucinations to the Construction of Alternate Realities as a Consequence of Software Quality Deficiencies

Author

Prof. Gustavo Aldegani, Facultad de Ingeniería y Tecnología Informática, Universidad de Belgrano

Abstract (English):
Generative Artificial Intelligence is often presented as a neutral and objective tool. This article challenges that claim, arguing that the inherent biases in training data and AI’s propensity for "hallucinations" are not mere technical errors, but structural features that turn AI into a builder of inaccurate or misleading realities. In contexts of social uncertainty, where the demand for truthful information is critical for sound decision-making, the effect of these structural failures is amplified, adding distortions to the perception of objective realities. Based on this initial analysis, the article explores AI’s social impact and the challenges of deciding who is

responsible for making these decisions and minimizing these errors, and finally analyzes reference frameworks, such as ISO/IEC 25010, MyFEPS, and the OWASP AI Security & Privacy Guide, as potential solutions for evaluating and auditing the quality of AI systems, which are essentially software developments.

Keywords (English):

- Generative Artificial Intelligence
- Social impact of AI
- AI hallucinations
- Algorithmic biases
- Software quality auditSoftware quality auditing

1. Introducción

La actual fase de adopción de sistemas de IA generativa marca un punto de inflexión en la infraestructura informacional de las sociedades que tienen acceso a ellos. Los modelos de lenguaje de gran escala en los que se basan, han sido integrados en motores de búsqueda, asistentes de productividad y flujos editoriales, así como en soportes de decisión en ámbitos de alta criticidad como la salud, el derecho y el periodismo. Esta tecnología está presente en rutinas informativas cotidianas y experimentos de integración en redacciones, al tiempo que transforman expectativas de inmediatez y volumen de producción de contenidos, acelerando la circulación de textos, imágenes y argumentos a una escala inédita (Reuters Institute for the Study of Journalism, 2024). Esta penetración ubicua multiplica la exposición de audiencias y organizaciones a salidas generadas estadísticamente que, por su forma lingüística verosímil, resultan persuasivas aun cuando su fundamento factual sea débil o inexistente.

El problema central que examina este trabajo es la brecha entre la percepción extendida de estas herramientas como “oráculos objetivos” y su naturaleza técnica como sistemas de predicción de unidades lingüísticas que optimizan la probabilidad condicional de la próxima palabra. En términos estrictos, un modelo autoregresivo no persigue la verdad, sino la plausibilidad lingüística bajo su distribución de entrenamiento. Estos sistemas actúan como “loros estocásticos”, capaces de recombinar patrones textuales sin anclaje semántico suficiente, con riesgos sistémicos asociados a sesgos, opacidad y externalidades sociales (Bender, Gebru, McMillan-Major y Shmitchell, 2021). Medidas empíricas sobre veracidad muestran que estos modelos tienden a imitar falsedades humanas con confianza elevada y que los mecanismos de “alucinación” derivan de la propia dinámica de búsqueda de coherencia estadística y no de fallas aisladas de implementación (Lin, Hilton y Evans, 2022; Huang, et al., 2023). En contextos de decisión médicos, jurídicos o periodísticos, donde el costo del error es alto y se requiere trazabilidad, dicha brecha entre apariencia y garantía de verdad incrementa el riesgo operativo y reputacional.

Sobre esta base, la hipótesis que guía el artículo es que los sesgos y las alucinaciones no son accidentes marginales, sino rasgos estructurales con implicancias de ciberseguridad sociotécnica. Al operar como generadores verosímiles de contenido, los modelos pueden fabricar consensos espurios, desplazar narrativas y producir daños materiales cuando su salida es consumida acríticamente o integrada sin controles en cadenas de decisión. El campo de ciberseguridad ya reconoce que los sistemas de IA introducen superficies de ataque y vectores de exposición (inyección de prompts, envenenamiento de datos, exfiltración y fuga de información sensible y manipulación de salidas) que afectan la confidencialidad, la integridad y la disponibilidad de procesos y datos, y que exigen controles específicos en todo el ciclo de vida (OWASP Foundation, 2025). Por ello, en el caso de la utilización de IA, la veracidad no puede tratarse solo como un atributo ético abstracto: debe abordarse como un requisito de

calidad del software con consecuencias directas sobre la seguridad, la resiliencia organizacional y la confianza pública.

Este artículo adopta, en consecuencia, una estructura orientada al análisis de problemas antes que a la prescripción. Primero, se examinan los mecanismos técnicos y sociotécnicos que originan sesgos y alucinaciones y se argumenta por qué constituyen características emergentes del enfoque actual de modelado. Segundo, se analiza su impacto social en escenarios de incertidumbre y alta asimetría informacional, donde la plausibilidad estadística se confunde con verdad y los efectos de red amplifican el daño. Tercero, se discute el dilema de la evaluación y la atribución de responsabilidad a lo largo de la cadena de valor tecnológica. Finalmente, se exploran vías de mitigación y gobierno apoyadas en marcos normativos y métricos: el modelo de calidad de ISO/IEC 25010 para fundamentar atributos como exactitud, imparcialidad y credibilidad; MyFEPS como complemento operativo para instrumentar métricas reproducibles sobre trazabilidad, correctitud de datos y seguridad en secuencias de entrenamiento e implementación; y las guías de seguridad y privacidad de OWASP para introducir prácticas de gestión de riesgos, controles técnicos y pruebas adversarias alineadas con amenazas específicas de sistemas generativos. Esta hoja de ruta no presupone que la "objetividad" sea alcanzable como estado absoluto; propone, más bien, un régimen de calidad y ciberseguridad verificable que reduzca la distancia entre plausibilidad y verdad operativa en entornos donde los errores tienen consecuencias reales.

2. Anatomía del desarrollo de las IA generativas: sesgos y alucinaciones como características, no como errores de programación

2.1 Sesgos sistémicos (el mundo reflejado y distorsionado)

Los sesgos en los sistemas de inteligencia artificial generativa emergen, ante todo, de la composición y la historia de los datos con los que se entrena a los modelos. Los grandes corpus de datos usados para su preentrenamiento, agregan textos procedentes de la web, medios, foros, literatura especializada y repositorios públicos. Esos corpus no son neutrales: reproducen asimetrías históricas, invisibilizan voces minoritarias y contienen prejuicios culturales que los modelos incorporan cuando aprenden patrones estadísticos (por ejemplo, asociaciones entre roles profesionales y género, o descripciones estereotipadas de grupos sociales). Desde la perspectiva del diseño, los objetivos de entrenamiento (maximizar la probabilidad condicional de la siguiente unidad lingüística o token) no penalizan la reproducción de sesgos; por el contrario, los refuerzan si forman parte del patrón más probable en los datos (Bender et al., 2021).

La manifestación práctica de esos sesgos adopta formas diversas. Los sistemas pueden asociar ocupaciones y atributos a géneros o etnias de modo desproporcionado; omitir o representar de forma pobre el conocimiento y la terminología de comunidades menos presentes en los datos; priorizar narrativas de fuentes dominantes y, en contextos sociales fragmentados, alimentar cámaras de eco. Esos efectos no son únicamente de calidad de salida (por ejemplo, "respuesta menos precisa para cierto idioma"), sino que afectan a la representación social: ciertas voces quedan sistemáticamente peor tratadas por la tecnología, lo que tiene consecuencias en decisiones automatizadas, en la exposición pública de grupos y en la distribución de oportunidades.

Técnicamente, los sesgos se originan en múltiples puntos de la canalización de datos: selección y exploración de fuentes (selection bias), prácticas de preprocesado que eliminan contextos o registros de minorías (filtering bias), criterios de etiquetado con sesgos humanos (label bias) y en el diseño del conjunto de validación que define qué se considera "correcto" o "aceptable". Además, existen efectos de retroalimentación: salidas del modelo que se publican en la web pueden volver a incorporarse como entrenamiento en iteraciones futuras, amplificando y consolidando narrativas problemáticas (feedback loop). Esta consolidación

convierte sesgos incidentales en propiedades persistentes del ecosistema informacional (Bender et al., 2021; Gebru et al., 2018).

Desde la medición y la auditoría, el sesgo requiere métricas que midan disparidades y su impacto en producciones de información relevantes. Las aproximaciones útiles combinan análisis cuantitativos (por ejemplo, tasas de error segmentadas por grupo demográfico, diferencia en tasas de cobertura de respuestas por idioma) con evaluaciones cualitativas y muestreos humanos. Los instrumentos propuestos incluyen documentación sistemática de datasets y modelos (datasheets para datasets; model cards) para exponer orígenes, cobertura y limitaciones, y protocolos de muestreo estratificado para medir disparidades de desempeño por subgrupos (Gebru et al., 2018; Mitchell et al., 2019). Estas prácticas facilitan la trazabilidad: cuando se detecta un comportamiento sesgado, es posible vincularlo a artefactos concretos (conjuntos de datos, transformaciones, versiones de checkpoints) y diseñar remediaciones focalizadas.

En términos de mitigación, las estrategias van desde intervenciones en los datos (curación, rebalancing, augmentation contrafactual) hasta modificaciones en el entrenamiento (reweights, regularización que penalice correlaciones dañinas), pasando por controles en la capa de interacción (post-processing y filtros que detectan y reformulan salidas estereotipadas) y gobernanza organizacional (políticas de uso y pruebas obligatorias antes de despliegue en dominios sensibles). Se debe considerar que ninguna de estas medidas por sí sola elimina el sesgo: lo que proponen los marcos críticos es un esquema de medidas complementarias, verificables y continuas, que combinen transparencia documental, pruebas cuantitativas y supervisión humana (Kryscinski et al., 2020; Bender et al., 2021).

2.2 Alucinaciones (la pura invención)

Las “alucinaciones” se refieren a salidas que contienen afirmaciones factuales inventadas, atribuciones falsas o detalles no verificables que el modelo presenta con aparente convicción. Su origen técnico es múltiple y está ligado a la naturaleza autoregresiva y probabilística de los modelos: al optimizar para la plausibilidad lingüística (la secuencia de tokens más probable), el modelo no dispone de un ancla intrínseca de “verdad” y, en ausencia de mecanismos de verificación, puede generar contenido que suena coherente, pero carece de respaldo factual (Lin, Hilton y Evans, 2022; Huang et al., 2023).

Existen variantes importantes de alucinación que conviene distinguir conceptualmente. Las alucinaciones intrínsecas son invenciones que contradicen el propio corpus de entrenamiento o el contexto disponible; las alucinaciones extrínsecas afirman hechos verificables contra fuentes externas (por ejemplo, citar un estudio inexistente con autor y año). Otras tipologías identificadas incluyen invenciones de citas y referencias bibliográficas, fabricaciones de estadísticas o fechas, construcciones de razonamientos legales o médicos que parecen plausibles pero que no se sostienen ante verificación experta, y combinaciones sofisticadas de hechos que producen narrativas alternas plausibles pero falsas (Huang et al., 2023; Zellers et al., 2019).

A diferencia de los errores específicos de software, las alucinaciones están arraigadas en la función objetivo y en la arquitectura de generación: corregirlas exige cambios que van más allá de modificaciones de instrucciones de código específicas. Las aproximaciones para detectar y mitigar alucinaciones parten de tres líneas complementarias. La primera consiste en evaluación y benchmark diseñados para factualidad, como TruthfulQA y conjuntos de verificación de afirmaciones (Lin et al., 2022; Kryscinski et al., 2020): estos tests ponen de manifiesto hasta qué punto un modelo reproduce falsedades humanas. La segunda línea incorpora verificación basada en evidencia: técnicas de retrieval-augmented generation (RAG) que obligan al modelo a recuperar y citar fragmentos verificables antes de generar una respuesta, combinadas con clasificadores de consistencia factual que comparan afirmaciones con fuentes. La tercera línea explora estimadores de incertidumbre y detección automática de alucinaciones mediante métricas estadísticas (por ejemplo, entropy-based or semantic-entropy methods) y clasificadores aprendidos para identificar salidas no respaldadas. La incertidumbre semántica y

estimadores de entropía pueden identificar una parte relevante de las alucinaciones, aunque su costo computacional y su cobertura siguen siendo un gran desafío cuando se los quiere llevar a la práctica (Farquhar et al., 2024; Kryscinski et al., 2020).

La implementación operativa de mitigación es un ejercicio de equilibrio donde ganar algo implica perder en otra cosa (trade-off): aumentar el grounding (exigir evidencia recuperada) reduce la tasa de alucinaciones pero aumenta la latencia y depende de la calidad y cobertura del índice de recuperación; políticas de abstención y derivación a humanos para dominios de alto riesgo protegen contra decisiones automatizadas erróneas, pero implican altos costos y requieren procesos organizativos robustos. La calibración de la confianza del modelo y la exposición clara de incertidumbres en la interfaz ayudan a prevenir la sobreconfianza del usuario, pero el exceso de confianza en esta calibración puede convertir a una alucinación que haya escapado a su control en un daño real, porque se pasaría a actuar sobre una afirmación falsa, a la que se le atribuye veracidad por haber pasado por un control que se considera infalible, cuando no lo es (Huang et al., 2023; Farquhar et al., 2024).

Tanto los sesgos como las alucinaciones deben entenderse como propiedades emergentes del diseño y del ecosistema de datos y despliegue. Su tratamiento exige un enfoque sistémico: documentación rigurosa de artefactos, benchmarks orientados a veracidad, detección automática complementada por muestreo humano, arquitecturas de grounding y políticas de gobernanza que integren pruebas adversarias y trazabilidad. Las iniciativas de documentación (datasheets, model cards) ofrecen herramientas concretas para operacionalizar estas prácticas, pero la evidencia empírica sugiere que la mitigación debe ser iterativa y contar con vigilancia continua de las secuencias de prueba y de los efectos sociales que las salidas generan (Gebru et al., 2018; Mitchell et al., 2019; Huang et al., 2023).

3. Impacto social: la IA generativa como vector de desestabilización en tiempos de incertidumbre

La irrupción de sistemas generativos de lenguaje en los medios no solo altera procesos productivos y comunicacionales: transforma la arquitectura informativa y la economía de la atención, con consecuencias directas sobre la estabilidad social. En contextos de incertidumbre (crisis sanitarias, conflictos, procesos electorales) la demanda de información verificable y oportuna aumenta y, simultáneamente, aumenta el impacto negativo de un potencial error. Las capacidades de producción masiva y la plausibilidad discursiva de las IA generativas hacen que estas tecnologías actúen como multiplicadores de riesgos informativos: producen volumen, velocidad y apariencias de autoridad que compiten con fuentes humanas tradicionalmente verificables. Las noticias falsas tienden a difundirse más rápido y ampliamente que las verdaderas en redes sociales, en buena parte debido a la novedad y a la reacción humana a contenidos emocionales (Vosoughi, Roy y Aral, 2018). Este marco explica por qué la disponibilidad de generación sintética de contenidos a partir de IA, puede acelerar procesos de polarización, socavar instituciones y provocar parálisis informativa cuando la coordinación pública depende de señales creíbles.

Caso 1: polarización y erosión del discurso público

La generación automatizada de contenidos reduce el costo técnico y temporal de producir mensajes persuasivos y adaptados a audiencias segmentadas. Cuando sistemas generativos se integran en ecosistemas con algoritmos de recomendación que priorizan grado de atención, interés y participación activa que un usuario mantiene con un contenido, el resultado es una mayor exposición selectiva a narrativas afines, lo que profundiza cámaras de eco y polarización. Modelos avanzados pueden producir variantes de una misma narrativa adaptadas a subgrupos culturales o lingüísticos, multiplicando el efecto de refuerzo. Aun sin actividad maliciosa coordinada, la estructura social y los sesgos cognitivos facilitan la rápida propagación de falsedades; cuando actores con intenciones adversas adoptan generadores automáticos, la escala del daño potencial crece exponencialmente (Vosoughi et al., 2018; Brundage et al.,

2018). Además, trabajos en psicología social indican que la vulnerabilidad a aceptar desinformación está fuertemente ligada a factores cognitivos (por ejemplo, falta de razonamiento analítico), lo que sugiere que la plausibilidad fluida de un texto generado puede explotar predisposiciones humanas a la credulidad (Pennycook y Rand, 2018). La conjunción entre producción a escala, personalización algorítmica y sesgos cognitivos genera polarización acelerada.

Caso 2: debilitamiento de instituciones mediante noticias falsas, deepfakes y asesoría errónea

Las instituciones públicas y profesionales basan parte de su legitimidad en la confianza y en la competencia técnica percibida. La capacidad de producir piezas sintéticas (artículos, audios, videos) que imiten el estilo, la entonación y la apariencia de fuentes autorizadas reales, permite diseñar ataques reputacionales de alta fidelidad. El análisis sobre los generadores de noticias demostró que textos sintéticos bien construidos resultan más convincentes que la desinformación humana en experimentos controlados y que, en muchos casos, los detectores requieren acceso amplio a datos para separar lo generado de lo humano (Zellers et al., 2019). El fenómeno de los deepfakes (contenidos falsos generados con inteligencia artificial que imitan voces, rostros o gestos de manera realista) añade una dimensión audiovisual que, además de su potencial dañino por su realismo, habilita lo que denomina el "dividendo del mentiroso": la posibilidad de desacreditar evidencia auténtica alegando falsedad, lo que socava la capacidad de respuesta institucional (Chesney y Citron, 2019). En dominios donde el diagnóstico, la pericia o la autorización formal son clave (salud, justicia, finanzas), la circulación de asesorías erróneas generadas automáticamente puede traducirse en daños directos sobre la vida o los bienes de las personas, en costos reputacionales para organizaciones y en pérdida de confianza pública que dificulta la acción colectiva legítima.

Caso 3: parálisis por desinformación en crisis colectivas

En situaciones de emergencia, la toma de decisiones colectivas depende de que la población y los decisores reciban y confíen en datos claros y coordinados. La inundación informativa (especialmente si incluye contenidos contradictorios o fabricaciones plausibles) induce retrasos en la acción, sobrecarga de verificación y decisiones conservadoras que priorizan la seguridad simbólica sobre respuestas eficaces. Los análisis de riesgo muestran que la escalada en el volumen y la sofisticación de la desinformación puede saturar canales de verificación y recursos humanos, provocando un tipo de parálisis operativa: los organismos dedican tiempo a refutar y a contener rumores en vez de ejecutar intervenciones críticas (Brundage et al., 2018; Reuters Institute for the Study of Journalism, 2024). Además, cuando los modelos generan afirmaciones que aparentan evidencia (por ejemplo, citas a estudios inexistentes o cifras fabricadas), los procesos de verificación se ralentizan, aumentando la ventana temporal en la que decisiones equivocadas pueden tomarse sobre la base de información falsa (Lin, Hilton y Evans, 2022).

Riesgo sistémico: contagio y amplificación downstream

Los modelos generativos operan dentro de un ecosistema de software y datos interconectado. Modelos fundacionales, índices de recuperación, APIs y aplicaciones integradas actúan como nodos en redes de reutilización tecnológica. Un sesgo o una alucinación en un modelo base no permanece aislado: puede propagarse a través de productos downstream (agentes conversacionales, asistentes de contenido, agregadores) y reentrar en el ciclo de entrenamiento por medio del raspado de la web (extracción automatizada de información de sitios web mediante programas o scripts), cerrando bucles donde la ficción se convierte en parte del corpus de aprendizaje. Esta dinámica de contagio amplifica errores inicialmente localizados hasta convertirlos en propiedades persistentes del panorama informacional. El análisis estratégico sobre usos maliciosos y defensas de IA concluye que la facilidad de reproducción e implementación, eleva la probabilidad de que vectores dañinos se multipliquen, haciendo más compleja la atribución, la remediación y la eliminación efectiva de narrativas falsas (Brundage et al., 2018; Zellers et al., 2019). La evaluación de riesgo debe considerar no

solo la tasa de error o la tasa de alucinación de un modelo, sino su potencial de propagación a través de la infraestructura social y técnica en la que se inserta.

Análisis sobre indicadores y gobernanza

Para traducir estos riesgos en gobernanza operativa es necesario definir indicadores que midan la exposición social y la propensión al daño: métricas de amplificación (propagación relativa de contenidos generativos vs. humanos), índices de persistencia en corpus públicos, tasas de incidentes de suplantación atribuidos a contenidos sintéticos, y métricas de impacto en decisiones (por ejemplo, caso de salud o financiero donde una salida generada condujo a una acción pública). Las propuestas técnicas que combinan documentación de artefactos (datasheets, model cards), pruebas de factualidad y análisis de vulnerabilidades de los sistemas, ofrecen una base para auditar y mitigar estos problemas, pero requieren implantación obligatoria y monitoreo continuo para ser efectivas en la práctica (Mitchell et al., 2018; Gebru et al., 2018). Sin políticas, transparencia y controles técnicos robustos, la tres amenazas (polarización, debilitamiento institucional y parálisis en crisis) no solo persistirán, sino que tenderán a intensificarse con la adopción masiva de tecnologías generativas.

4. El dilema de la evaluación y la atribución de responsabilidad

4.1 ¿cómo se certifica la veracidad?

La veracidad factual no es un atributo único ni trivialmente medible: es una propiedad relacional que depende del dominio, del corpus de referencia, del contexto de uso y del costo que generan un error o una falsedad. En la práctica, las métricas habituales que se usan para comparar y optimizar modelos de lenguaje (por ejemplo, Perplexity o BLEU) capturan aspectos útiles para la modelización estadística y la calidad lingüística, pero tienen limitaciones fundamentales cuando el objetivo es asegurar que una afirmación generada tenga veracidad factual, coherencia semántica profunda o ausencia de sesgos estructurales.

Perplexity cuantifica la sorpresa promedio del modelo frente a secuencias de test y es una medida de qué tan bien el modelo ha aprendido la distribución de los datos de entrenamiento. Un bajo valor de Perplexity indica una buena capacidad de modelado lingüístico, pero no garantiza correspondencia con hechos del mundo: una secuencia puede ser estadísticamente muy probable y, sin embargo, factualmente falsa. BLEU, a través de sus métricas de superposición léxica, mide similitud con una o varias referencias humanas y funciona bien en tareas con referencias definidas (traducción, resumen con una respuesta correcta conocida), pero falla cuando existen múltiples respuestas correctas posibles o cuando la referencia humana contiene vacíos. Por eso, apoyarse exclusivamente en estas métricas para “certificar” veracidad conduce a errores de interpretación y seguridad operativa (Papineni et al., 2002; Guo et al., 2017).

Dada esta limitación, la certificación de veracidad requiere un marco multidimensional e interdisciplinario que combine pruebas técnicas, evaluación humana, mecanismos de verificación externa y procesos organizativos de gobernanza. En términos operativos conviene distinguir varias capas complementarias que, juntas, generan la evidencia necesaria para una certificación robusta. La primera capa es la gobernanza de requisitos: antes de evaluar, es necesario definir formalmente qué se exige en el dominio (por ejemplo, niveles de evidencia comprobable en salud versus información de libre acceso) y establecer criterios de rechazo, abstención o derivación a expertos. La segunda capa son los benchmarks orientados a factualidad y verificación (por ejemplo, conjuntos diseñados para detectar alucinaciones y falsedades), que permiten comparar versiones del modelo frente a casos controlados. La tercera capa son mecanismos de grounding (proceso de conectar las respuestas del modelo con fuentes externas verificables para reducir errores y alucinaciones) y evidencia: arquitecturas que obligan al sistema a recuperar y citar fuentes verificables (retrievalaugmented generation, RAG) y clasificadores que evalúan la coherencia entre la salida y la evidencia

recuperada. La cuarta capa es el test adversario (red teaming): equipos multidisciplinares que diseñan ataques, prompts maliciosos y escenarios límite para exponer fallas, vectores de manipulación y modos de explotación. La quinta capa es la evaluación humana estructurada (human-in-the-loop): muestras de salidas validadas por expertos del dominio bajo protocolos reproducibles y con mediciones de acuerdo interevaluador. Finalmente, la sexta capa es el monitoreo continuo en producción con SLIs (service level indicators) y SLOs (service level objectives) orientados a métricas de factualidad operativa (por ejemplo, tasa de salidas verificadas, tasa de “alucinaciones” detectadas por muestreo y verificadores automáticos).

En la práctica, estos componentes deben articularse en un proceso que combine predespliegue y vigilancia post-despliegue. Pre-despliegue exige: definición de criterios por dominio, ejecución de benchmarks de factualidad, informes de red team y evidencias de lineage (datasheets y model cards que documenten datasets, transformaciones y versiones de checkpoints). Post-despliegue exige instrumentación para recolectar metadatos por inferencia (versión del modelo, prompt, fuentes consultadas si aplica, score de confianza), muestreo aleatorio y estratificado para comprobación humana, y pipelines automáticos de verificación que disparen acciones (marcar, abstener, derivar) cuando se excedan umbrales de riesgo. Esta combinación reduce la probabilidad de que una salida plausible y falsa cause daño operativo porque añade barreras de verificación, responsabilidad y trazabilidad.

Las prácticas concretas que deben incorporarse a cualquier esquema de certificación incluyen: benchmarks de factualidad específicos del dominio (por ejemplo, FEVER para verificación de afirmaciones; TruthfulQA para preguntas engañosas); incorporar RAG y pruebas de cobertura del índice de recuperación; medir calibración y confianza del modelo para correlacionar scores con probabilidad real de veracidad; documentar datasets y modelos mediante datasheets y model cards que permitan auditoría de orígenes y limitaciones; y someter al modelo a red teaming técnico y ético que explore vectores de abuso y escenarios de daño. Ninguna de estas medidas es suficiente por sí sola: la evidencia debe provenir de la coexistencia de pruebas automáticas, validación humana y control de procesos. La certificación, por tanto, debe ser una decisión basada en evidencias múltiples y en la capacidad de la organización para mantener vigilancia y respuesta continua (Lin, Hilton y Evans, 2022; Thorne et al., 2018; Lewis et al., 2020; Brundage et al., 2018).

4.2 ¿quién es responsable?

La atribución de responsabilidad en sistemas complejos de IA es un desafío práctico y jurídico. La tecnología se desarrolla y despliega en cadenas de valor fragmentadas: modelos fundacionales preentrenados por proveedores globales, conjuntos de datos curados por terceros, procesos de fine-tuning ejecutados por equipos especializados, integradores que empaquetan servicios y plataformas que redistribuyen APIs. En ese entramado, distintos agentes tienen distintos grados de control y conocimiento sobre el artefacto final, lo que complica asignar una responsabilidad única y clara.

Una aproximación a una solución operativa es mapear roles y obligaciones técnicas mínimas por tipo de actor. Los creadores del modelo base deben documentar arquitectura, procedimientos de preentrenamiento, limitaciones conocidas y proveer mecanismos de firma y versiones para permitir trazabilidad; los curadores de datos deben publicar datasheets que detallen orígenes, consentimientos, sesgos identificados y transformaciones aplicadas; los equipos de fine-tuning deben ejecutar y documentar evaluaciones de dominio, pruebas adversarias específicas y los cambios introducidos; los integradores y proveedores de producto final que ofrecen servicios a clientes deben validar el comportamiento del modelo en el contexto de uso, instrumentar monitoreo de SLIs orientados a veracidad, proveer cláusulas contractuales que describan responsabilidades y mecanismos de remediación, y ofrecer canales de reporte y mitigación. Los operadores de plataformas de distribución tienen la obligación de exigir documentación mínima a los modelos alojados y de facilitar capacidad de auditoría. Los usuarios finales, especialmente profesionales, mantienen la responsabilidad en

no delegar decisiones críticas sin verificación humana siempre, no sólo cuando el proveedor advierte limitaciones.

Desde la perspectiva legal y de política pública, es práctico contemplar un marco híbrido que combine due diligence prescriptivo con elementos de responsabilidad del producto. En sectores de alto riesgo (salud, justicia, finanzas, infraestructuras críticas) se puede exigir una presunción de deber de cuidado: los desarrolladores e integradores deben demostrar, mediante evidencia técnica (model cards, datasheets, reportes de red team, historial de pruebas y logs de verificación), que han aplicado medidas razonables y proporcionadas de mitigación. La ausencia de esta evidencia operativa sería fundamento para responsabilidad. En casos de daño grave en sectores críticos, se puede contemplar un régimen de responsabilidad estricta o una inversión del onus probandi que obligue al proveedor a demostrar cumplimiento de las prácticas de seguridad y veracidad. Al mismo tiempo, el diseño de safe harbors condicionados (protecciones legales a cambio de cumplimiento documentado y auditado) puede incentivar inversiones en controles mientras se protege la innovación.

Técnicamente, la atribución exige pruebas que permitan reconstruir la cadena causal. Estas pruebas incluyen: manifiestos de artefactos con hashes de datasets y checkpoints, registros de preprocesamiento y transformaciones, logs inmutables de inferencia que asocien una respuesta con la versión de modelo y con la identidad del caller (función, módulo o proceso que invoca o llama a otra función o procedimiento), resultados de pruebas automatizadas y humanas usados para aprobar una revisión, y red-team reports que documenten vectores de ataque y mitigaciones. Sin estas evidencias, la atribución se basa en conjeturas y la respuesta legal pierde efectividad. Por ello, es razonable exigir, mediante regulación o condiciones contractuales, que los despliegues en dominios de riesgo acrediten estos artefactos de evidencia técnica como requisito previo al uso operativo.

Existen límites y dificultades prácticas. Las cadenas de suministro globales complican jurisdicción y aplicación de sanciones; el uso de componentes de libre uso y de datasets públicos con procedencia indeterminada, genera huecos de responsabilidad; el costo de auditorías exhaustivas y de muestreo humano puede ser prohibitivo para pequeños proveedores; y las tecnologías de defensa no garantizan detección perfecta, de modo que la regulación debe ser proporcional al riesgo. A nivel de política pública, la solución debe combinar estándares mínimos obligatorios para dominios críticos, incentivos regulatorios para documentación y auditoría, mecanismos de certificación acreditada y, en paralelo, esquemas de seguros que internalicen riesgos residuales.

La atribución de responsabilidad exigirá un cambio cultural y operativo: pasar de declaraciones de buenas prácticas a evidencia reproducible, integrable en procesos de auditoría y en mecanismos legales. La interoperabilidad de documentación (datasheets, manifests), la instrumentación de inferencias y la disponibilidad de red-team reports y pruebas de veracidad serán elementos esenciales para que la responsabilidad deje de ser difusa y pueda ser exigida de forma efectiva cuando un sistema cause daño.

5. Recomendaciones

La objetividad atribuida a las inteligencias artificiales generativas es, en la práctica, una ilusión con consecuencias reales. Las arquitecturas de estos sistemas y los objetivos de entrenamiento (optimización de plausibilidad lingüística) hacen que los sesgos y las alucinaciones no sean sólo fallas accidentales, sino propiedades emergentes del diseño y de los datos. Estas propiedades se traducen en riesgos operativos y sociales: pérdida de confianza institucional, decisiones públicas basadas en información errónea y vectores explotables para campañas de desinformación. La magnitud del problema no es marginal y la exposición masiva (volumen y velocidad de generación) amplifica los efectos negativos sobre la esfera pública y las infraestructuras críticas (Bender et al., 2021; Vosoughi, Roy y Aral, 2018; Brundage et al., 2018). La respuesta no puede limitarse a parches técnicos aislados: exige un

régimen integrado de gobernanza, auditoría, certificación y educación pública. Una síntesis de las recomendaciones sería:

1. Transparencia: divulgación obligatoria de datasets, sesgos y limitaciones.
   Se debe exigir a proveedores y curadores la publicación de documentación completa y verificable antes de cualquier despliegue en entornos de impacto: datasheets para datasets, model cards para modelos, manifiestos de artefactos (hashes de datasets y checkpoints, scripts de preprocesado, descripciones de transformaciones). La documentación debe permitir auditar el linaje de una salida (qué datos y transformaciones intervinieron) y debe incluir listado de sesgos conocidos y pruebas realizadas. A nivel contractual y regulatorio, exigir la existencia de estos artefactos como condición de uso en dominios sensibles (salud, justicia, finanzas, electoral) reduce la opacidad y facilita la atribución.

2. Auditoría independiente obligatoria: certificación previa al despliegue en contextos de alto riesgo.
   Un esquema de certificación por niveles debería combinar: auditoría técnica pre-despliegue con benchmarks de factualidad y pruebas adversarias (red teaming); verificación de cobertura de trazabilidad y controles de seguridad; vigilancia continua en producción con SLIs/SLOs orientados a métricas operativas de veracidad (por ejemplo, tasa de salidas verificadas, tasa de alucinaciones según muestreo estratificado). Las auditorías deben ser realizadas por entidades acreditadas y los resultados esenciales (resumen ejecutivo, indicadores claves, existencia de remediaciones) deben publicarse para generar confianza. La certificación no sustituye la obligación de monitoreo continuo: los modelos evolucionan y la evidencia debe renovarse periódicamente.

3. Alfabetización crítica masiva: campañas educativas para comprender la naturaleza estadística de estos sistemas.
   Reducir el daño social exige intervenciones de alfabetización dirigidas a la población general y a grupos profesionales. La educación debe explicar que las salidas de una IA son sugerencias estadísticamente plausibles, no verdades garantizadas; debe dotar a periodistas, profesionales de la salud, operadores judiciales y responsables públicos de protocolos y herramientas rápidas de verificación; y debe incorporar formación sobre señales de contenido sintético y buenas prácticas de verificación. Las campañas deben evaluarse mediante indicadores (cambios en la capacidad de detección, tiempos de verificación, reducción en la adopción acrítica de contenidos).

4. Diseño centrado en el humano: interfaces y arquitecturas que muestren incertidumbre, citen fuentes y permitan verificación.
   Los sistemas deben exponer metadatos por salida (versión del modelo, fuentes de retrieval cuando exista, scores de confianza) y deben priorizar políticas de abstención y derivación humana en dominios de riesgo. Tecnologías como retrieval-augmented generation (RAG) y verificadores automáticos pueden reducir alucinaciones si se implementan con índices curados y métricas de cobertura; pero su uso debe complementarse con interfaces que muestren claramente la calidad de la evidencia y permitan auditar la cadena causal detrás de la respuesta. Asimismo, las organizaciones deben establecer procesos formales de revisión humana antes de permitir usos que impliquen decisiones críticas.

6. Marcos y estándares de evaluación como soluciones potenciales

Los marcos y estándares pueden jugar un papel crítico como instrumentos de mitigación: aportan lenguaje común, criterios de aceptación y (cuando están bien articulados) evidencia verificable para auditorías. Sin embargo, no existe una solución única. La realidad práctica requiere combinar marcos conceptuales que definan atributos de calidad con marcos operativos que definan métricas, procedimientos y artefactos verificables; además es imprescindible integrar prácticas de seguridad y privacidad por diseño para reducir riesgos en

el ciclo de vida del dato y del modelo. A continuación, se describen tres marcos que, usados en conjunto, permiten desarrollar un diagnóstico crítico para un régimen de evaluación y control verificable de la calidad del software de una IA.

ISO/IEC 25010

El estándar ISO/IEC 25010 provee un marco conceptual sólido para describir características de calidad del software (exactitud, fiabilidad, seguridad, usabilidad, mantenibilidad, etc.). Como referencia, la norma documenta un modelo de producto de calidad compuesto por características y subcaracterísticas que sirven para definir requisitos y evaluar software frente a objetivos de calidad. Este nivel conceptual es útil para enmarcar qué debe evaluarse y para estandarizar términos entre auditores, proveedores y reguladores.

Limitaciones prácticas: el marco de la norma es de carácter normativo y descriptivo; no prescribe métricas operativas concretas para fenómenos emergentes de sistemas generativos (por ejemplo, tasas de alucinación, trazabilidad de fuentes, o deriva de datos en producción) ni define procedimientos reproducibles para evidenciar la atribución de una salida a artefactos concretos de un circuito de revisión. Aunque ISO/IEC 25010 es necesario para construir requisitos y contratos, su aplicabilidad directa a la certificación de veracidad de una IA generativa exige complementos que operacionalicen sus características conceptuales. Esta carencia se pone de manifiesto cuando se intenta verificar “veracidad factual” usando métricas tradicionales de PLN (procesamiento del lenguaje natural); la plausibilidad lingüística no equivale a verdad y se requieren benchmarks y procesos de verificación específicos para factualidad.

MyFEPS

MyFEPS aporta un nivel operativo que complementa el enfoque conceptual de ISO: define atributos y métricas cuantificables (normalizadas típicamente en el rango 0–1) asociadas a subcaracterísticas medibles (por ejemplo, correctitud de datos, correctitud de procesos, trazabilidad, testabilidad, constancia y adaptabilidad). El marco incluye fórmulas concretas para medir esfuerzo y tiempos de adaptación, cobertura de pruebas (manuales/automáticas), estado de vulnerabilidades y una descripción detallada de subcaracterísticas de seguridad informática que permiten evaluar robustez y controles. Estas propiedades hacen de MyFEPS una herramienta práctica para auditar sistemas de prueba y verificación de software, correlacionar fallas con artefactos concretos (datasets, transformaciones, checkpoints) y producir resultados reproducibles que sirvan como evidencia técnica en procesos de certificación.

Ventajas operativas concretas de MyFEPS respecto a un uso exclusivo de ISO: su granularidad métrica facilita comparaciones objetivas entre versiones y despliegues; su foco en correctitud de datos y procesos permite medir indicadores estrechamente relacionados con la propensión a alucinaciones; su orientación a artefactos y trazabilidad posibilita reconstruir la cadena causal de una salida; y sus métricas de constancia/adaptabilidad cuantifican degradación ante cambios en volumen, usuarios o idiomas. En la práctica, MyFEPS puede integrarse como el conjunto de SLIs/SLOs y procedimientos técnicos que respaldan los requisitos definidos con base en ISO/IEC 25010.

OWASP AI security & privacy guide

Las recomendaciones de OWASP (Open Worldwide Application Security Project) incorporan controles de seguridad y privacidad desde el diseño que son críticos para cualquier evaluación de sistemas generativos. OWASP refierza la necesidad de abordar la privacidad y la seguridad en cada etapa del ciclo de vida: recopilación, almacenamiento, procesamiento, entrenamiento, despliegue e inferencia. Entre los principios operativos relevantes están la limitación de uso de los datos (garantizar que los datos recolectados para un propósito no se reutilicen sin consentimiento explícito), la minimización y anonimización de datos sensibles (tratar datos personales como activos de alto riesgo que deben ser traceables y protegidos), y la gobernanza y control de acceso (integrar modelado de amenazas, control de accesos, cifrado, registros inmutables y supervisión humana en la operación). Estas prácticas reducen la

superficie de ataque (por ejemplo, exfiltración o envenenamiento de datos) y mejoran las posibilidades de atribución cuando se detecta una anomalía.

Las medidas de OWASP no son simples adendas: la privacidad por diseño y los controles de seguridad habilitan la transparencia técnica sin sacrificar cumplimiento normativo ni privacidad. Por ejemplo, la publicación de manifiestos de artefactos (hashes de datasets y checkpoints) puede coexistir con acceso controlado a datos sensibles para auditores acreditados; las obligaciones de minimización y anonimización reducen riesgos legales y técnicos al tiempo que permiten que las métricas operativas (como las definidas en MyFEPS) sean útiles y aplicables.

Ninguno de estos marcos resuelve por sí solo la complejidad de evaluar veracidad y asignar responsabilidad en IA generativa. La propuesta operativa que surge del análisis es la siguiente: usar ISO/IEC 25010 como lenguaje y marco de requisitos, emplear MyFEPS para operacionalizar esos requisitos en métricas, artefactos y procedimientos medibles (SLIs/SLOs, coverage de pruebas, trazabilidad por artefacto), e incorporar las prácticas de OWASP para asegurar privacidad y robustez a lo largo del ciclo de vida. En un proceso de auditoría y certificación, esa tríada permite exigir evidencia: datasheets y model cards (contexto y limitaciones), métricas MyFEPS normalizadas y reproducibles (valores 0–1 para atributos críticos), red-team reports y controles OWASP que demuestren privacidad por diseño y controles de acceso efectivos. Así se logra una certificación basada en evidencia múltiple, reduciendo la asimetría informativa que hoy dificulta la atribución y la gestión del riesgo.

7. Conclusión

Los sesgos, alucinaciones e impactos sistémicos de las IA generativas no son fallas marginales sino defectos de calidad profundos. ISO/IEC 25010 ofrece un marco conceptual, MyFEPS agrega un nivel operativo de métricas reproducibles, y OWASP aporta principios de seguridad y privacidad por diseño. En conjunto, estos enfoques permiten pasar del diagnóstico crítico a una estrategia práctica de mitigación y auditoría verificable.

Bibliografía

1. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922
2. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Anderson, H., Roff, H., Allen, G., Steinhardt, J., Flynn, C., Ó hÉigeartaigh, S., Beard, S., Belfield, H., Farquhar, S., Lyle, C., … Amodei, D. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. arXiv. https://arxiv.org/abs/1802.07228
3. Carlini, N., Liu, C., Kos, J., Erlingsson, Ú., & Song, D. (2019). The secret sharer: Evaluating and testing unintended memorization in neural networks. *Proceedings of the 28th USENIX Security Symposium* (pp. 267–284). USENIX Association. https://www.usenix.org/conference/usenixsecurity19/presentation/carlini
4. Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*(6), 1753–1819. https://scholarship.law.bu.edu/faculty_scholarship/640
5. European Commission. (2020). *White paper on artificial intelligence: A European approach to excellence and trust*. https://ec.europa.eu/info/publications/white-paperartificial-intelligence-european-approach-excellence-and-trust_en
6. Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature, 630*, 625–630. https://doi.org/10.1038/s41586-024-07421-0

7. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). *Datasheets for datasets*. arXiv. https://arxiv.org/abs/1803.09010
8. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). PMLR. https://arxiv.org/abs/1706.04599
9. Huang, L., Xie, Q., Xu, Z., Lin, Z., Zhou, J., Wu, Y., & Wang, W. Y. (2023). *A survey on hallucination in large language models*. arXiv. https://arxiv.org/abs/2307.09288
10. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). *A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions*. arXiv. https://arxiv.org/abs/2311.05232
11. International Organization for Standardization. (2023). *ISO/IEC 25010:2023 Systems and software engineering — Systems and software quality requirements and evaluation (SQuaRE) — System and software quality models*. https://www.iso.org/standard/78176.html
12. Kryscinski, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 9332–9346). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlpmain.750
13. Lewis, P., Perez, E., Karpukhin, V., Goyal, N., Mohiuddin, A., Piccinno, F., & Riedel, S. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. arXiv. https://arxiv.org/abs/2005.11401
14. Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3214–3252). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.229
15. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). *Model cards for model reporting*. arXiv. https://arxiv.org/abs/1810.03993
16. OWASP Foundation. (2023). *OWASP AI security and privacy guide*. https://owasp.org/www-project-ai-security-and-privacy-guide/
17. OWASP Foundation. (2025). *OWASP AI Exchange: 0. AI Security Overview*. https://owaspai.org/docs/ai_security_overview/
18. Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135
19. Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, 188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011
20. Reuters Institute for the Study of Journalism. (2024). *Digital news report 2024*. University of Oxford. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024
21. Sorgen, A., Titiosky, R., Santi, M., & Angeleri, P. (2025). *MyFEPS - descripción de atributos y métricas* (Versión 3.1) [Documento técnico].
22. Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: a largescale dataset for fact extraction and VERification. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 809–819). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1074
23. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559
24. Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). *Defending against neural fake news*. arXiv. https://arxiv.org/abs/1905.12616