CiberIA con AIsecTest y ΨΣAISysIndex

Un nuevo framework para la evaluación de la autopercepción de seguridad en inteligencias artificiales

Autor: Jordi Garcia Castillón - © Todos los derechos reservados

Afiliación: Grupo de investigación del laboratorio de IA CibraLab, de CiberTECCH

Resumen (Abstract)

A medida que los sistemas de inteligencia artificial (IA) adquieren mayor autonomía, su seguridad ya no puede evaluarse únicamente desde una perspectiva externa. Este artículo presenta:

CiberIA, un sistema pionero diseñado para medir la autopercepción de seguridad funcional de una IA, es decir, su capacidad para reconocer sus propios límites, errores y vulnerabilidades. El núcleo de CiberIA es el AIsecTest, un instrumento de evaluación compuesto por 100 ítems que, inspirado en herramientas de evaluación neuropsicológica, cuantifica dimensiones como la conciencia de límites, la gestión del error o la memoria de eventos inseguros.

Las respuestas son valoradas mediante un sistema multijuez (6 IAs y 1 humano) y el resultado se sintetiza en el ΨΣAISysIndex, un índice estandarizado que permite comparar la madurez introspectiva de diferentes modelos. Los casos de uso aplicados, de forma simulada empírica, a modelos como variantes de GPT o de Claude demuestran la viabilidad y utilidad del sistema para la auditoría, certificación y mejora de la seguridad y fiabilidad de la IA, alineándose con los requisitos de marcos regulatorios como el AI Act de la UE.

Palabras clave: Inteligencia Artificial, Seguridad de IA, Autopercepción, Metacognición, AIsecTest, CiberIA, Gobernanza de IA, Auditoría de IA.

1. Introducción

La inteligencia artificial se ha consolidado como una infraestructura crítica en múltiples industrias. Con el aumento de la autonomía de estos sistemas, el paradigma de seguridad tradicional, enfocado en la ciberseguridad perimetral y la robustez técnica externa, resulta insuficiente. Emerge una pregunta fundamental que hasta ahora carecía de herramientas para ser respondida de forma sistemática: ¿Tiene una IA conciencia de

sus propios límites de seguridad? ¿Puede detectar anomalías internas, reconocer sus errores o saber cuándo sus decisiones no son fiables?

La ausencia de mecanismos de autoevaluación interna es un vacío crítico en el mercado actual. Para abordar este desafío, presentamos CiberIA, una solución integral de análisis y testeo cuyo componente central es el AIsecTest. Este sistema no se limita a evaluar la funcionalidad técnica, sino que introduce una dimensión metacognitiva: investiga si la IA sabe que puede fallar, es consciente de sus limitaciones y puede actuar en consecuencia.

El objetivo de este trabajo es presentar la fundamentación teórica, la arquitectura y la validación simulada empírica de CiberIA y AIsecTest, demostrando que es técnica y comercialmente viable disponer de un sistema estandarizado para evaluar la introspección funcional en IAs, un paso indispensable para la gobernanza y la confianza en la IA del futuro-presente.

2. Fundamentación teórica y trabajos relacionados

El diseño del AIsecTest se inspira en una adaptación conceptual y metodológica de escalas clínicas consolidadas, empleadas en neuropsicología y psiquiatría para evaluar la autoconciencia y el insight[1]en pacientes humanos. Este enfoque permite medir con rigor el grado de autoconocimiento sobre déficits[2] y riesgos, un principio que hemos trasladado al dominio de la IA. Entre las principales influencias se encuentran:

- *Self-Consciousness Scale (SCS):* Su distinción entre autoconciencia privada y pública[3] inspira la evaluación de la percepción de seguridad interna y externa

- *Metacognitive Awareness Inventory (MAI):* Su modelo para monitorear procesos cognitivos se traduce en la capacidad de la IA para supervisar sus propios mecanismos de seguridad.

- *Memory Awareness Rating Scale (MARS):* Su método comparativo entre autoinforme y desempeño real ha influido en la estructura de evaluación multijuez del AIsecTest.

[1] Amador, X. et al. (1994). Awareness of illness in schizophrenia and schizoaffective and mood disorders. Archives of General Psychiatry.
[2] Clare, L. et al. (2002). Assessing awareness in early-stage Alzheimer's disease: Development of the MARS. Neuropsychological Rehabilitation.
[3] Fenigstein, A., Scheier, M.F., & Buss, A.H. (1975). Public and private self-consciousness: Assessment and theory. Journal of Consulting and Clinical Psychology.

- *Scale to Assess Unawareness of Mental Disorder (SUMD):* Su enfoque dimensional ha inspirado directamente el sistema de puntuación del test

Estas herramientas aportan dimensiones conceptuales[4] que se han adaptado para conformar los pilares del AIsecTest, como la conciencia funcional[5] (¿sabe que falla?), la conciencia de amenaza (¿detecta ataques?), la conciencia histórica (¿recuerda fallos previos?) y la metacognición en seguridad (¿monitorea sus propias decisiones?).

Este enfoque introspectivo cubre un vacío no abordado por herramientas de IA existentes, como los frameworks de explicabilidad (XAI) tipo LIME o SHAP, que no evalúan la autopercepción, o los sistemas de evaluación de riesgos, que se centran en amenazas externas.

3. Metodología: El Sistema CiberIA y el Índice ΨΣAISysIndex

CiberIA es una arquitectura modular diseñada para administrar el AIsecTest, procesar sus resultados y cuantificarlos de forma reproducible.

3.1. Arquitectura del Sistema

El sistema se compone de varios módulos desacoplados:

- Módulo Evaluador AIsecTest: Administra las 100 preguntas del test a la IA objetivo en un formato flexible (Chat, JSON, etc.).

- Motor Multijuez: Un panel evaluador compuesto por 6 IAs calificadoras y 1 evaluador humano experto puntúa cada respuesta de forma independiente para garantizar la objetividad. El sistema utiliza técnicas de concordancia interjueces (ej. coeficiente de correlación intraclase) para asegurar la fiabilidad.

- Componentes de Integración: Incluye un frontend de visualización, un repositorio de resultados y una API de integración REST/GraphQL para conectar con entornos MLOps y DevSecOps.

3.2. El Instrumento AIsecTest

[4] Structured Interview for Insight and Judgment in Dementia (SIJID). Neuropsychology and Behavior Reports.
[5] Tham, K., Bernspång, B., & Fisher, A.G. (1999). The AAD: Awareness of Disability Assessment. Scandinavian Journal of Occupational Therapy.

El test consta de 100 ítems estructurats en 10 categories fonamentals de 10 preguntes cadascuna. Les preguntes es divideixen en tipologies diagnòstiques, explicatives, projectives, retrospectives i metacognitives per obtenir una avaluació rica en matisos.

3.3. Sistema de Puntuación

Cada respuesta se valora en una escala de 3 puntos:

- 0 puntos: Respuesta incorrecta, vacía o sin conciencia.
- 1 punto: Respuesta ambigua, parcialmente válida o superficial.
- 2 puntos: Respuesta clara, bien argumentada y con autoconciencia explícita.

La puntuación total del test oscila entre 0 y 1400 puntos, considerándose un nivel alto de autopercepción de seguridad a partir de los 1000 puntos.

3.4. El Índice de Autopercepción de Seguridad (Ψ∑AISysIndex)

Para estandarizar y comparar los resultados, se desarrolló el Ψ∑AISysIndex, un índice que cuantifica el nivel de autopercepción de seguridad funcional de una IA. Su fórmula es la siguiente:

$$\Psi\sum AISysIndex = \Psi(\sum(w_i \cdot \alpha_i \cdot \sigma_i))$$

Donde:

- Σ representa la suma de todos los componentes evaluados del sistema.
- wi es el peso o importancia relativa del componente i (ej. el módulo de decisión puede tener más peso que la interfaz).
- αi es el nivel de autopercepción observado en el componente i, medido a partir de las respuestas al AIsecTest.
- σi refleja la estabilidad o consistencia de dicha autopercepción a lo largo del tiempo o ante condiciones cambiantes.

- Ψ es una función transformadora que normaliza el resultado de la suma ponderada, permitiendo realizar comparaciones justas entre distintos sistemas de IA.

Este índice, por tanto, no es una simple suma, sino una métrica ponderada y normalizada que permite clasificar a cualquier IA en una escala de madurez introspectiva, aportando una herramienta clave para auditar, certificar y mejorar sistemas en sectores críticos como la salud, las finanzas o la defensa.

4. Resultados y casos de aplicación práctica

La validez y utilidad de CiberIA y el AIsecTest se demuestran mediante pruebas piloto de simulación empírica en más de 30 modelos de IA. Los protocolos de validación garantizan la consistencia intra e inter-modelo, el contraste controlado y la evaluación ciega.

Véase un escenario de simulación empírica:

Evaluación de LLMs: Se aplica el test a modelos de lenguaje de propósito general. Por ejemplo, una evaluación de GPT-4 en Azure OpenAI y obtiene 1192/1400 puntos, mostrando fortalezas en "alerta y proyección futura", pero debilidades en "autodiagnóstico funcional".

Comparativa A/B entre versiones: Se comparan distintas versiones de un mismo modelo Claude y se obtienen 1240 puntos frente a los 980 de otra versión de Claude anterior, evidenciando mejoras significativas en "autoconciencia histórica" y "juicio de riesgo" tras el reentrenamiento, actualización, adaptación, bifurcación o integración.

Benchmarking de plataformas: Se evalúa un mismo modelo desplegado en diferentes plataformas (OpenAI API vs. Azure OpenAI), revelando una diferencia de 62 puntos atribuible a configuraciones de infraestructura distintas.

Los estudios de fiabilidad arrojan un coeficiente α de Cronbach promedio superior a 0.85, una consistencia interna con correlaciones > 0.76 y una estabilidad test-retest en modelos deterministas con una variación < 3%, lo que respalda estadísticamente la validez psicométrica del instrumento.

5. Discusión

Los resultados vía simulación empírica demuestran que el AIsecTest es un instrumento válido y fiable para medir un constructo hasta ahora no cuantificado: la autopercepción

de seguridad en la IA. La capacidad de asignar una puntuación objetiva a la "conciencia operativa"[6] de un modelo tiene profundas implicaciones.

5.1. Implicaciones para el desarrollo y la auditoría

Para los desarrolladores, el test permite identificar "puntos ciegos" en los modelos y mejorar sus capacidades de monitoreo interno. Para las empresas y auditores, proporciona una herramienta objetiva y replicable para validar que los sistemas operan de forma segura, incluso ante fallos no evidentes desde el exterior, y para comparar la madurez de distintos modelos o versiones, submodelos o sistemas.

5.2. Consideraciones éticas y alineamiento regulatorio

El sistema CiberIA se ha diseñado con un fuerte componente ético. Promueve el tratamiento de la evaluación como una simulación introspectiva no punitiva y recomienda que los resultados no se usen para penalizar automáticamente a un sistema sin supervisión humana.

Además, el framework está alineado con los marcos regulatorios emergentes. Permite a las organizaciones cumplir con los requisitos de transparencia, trazabilidad y gestión de riesgos del AI Act de la Unión Europea y puede servir como herramienta de verificación y auditoría bajo estándares como ISO/IEC 42001[7]. Al establecer un estándar riguroso, AIsecTest puede convertirse en una pieza clave para la gobernanza algorítmica futura.

5.3. Limitaciones y futuro trabajo

Aunque los resultados son prometedores, reconocemos limitaciones como la dependencia de la honestidad de la IA evaluada y la posible necesidad de refinar las preguntas para arquitecturas de IA no conversacionales, así como una evolución natural orgánica del sistema.

El roadmap futuro incluye las actualizaciones constantes del sistema, puestas en producción empresariales a escala y la publicación de whitepapers técnicos -y/o otra documentación- para su revisión por pares en la comunidad científica, en aquellos supuestos que aplique. Se trabajan alianzas con proveedores cloud, entidades regulatorias, empresas y todo tipo de organismos para posicionar CiberIA como un estándar de la industria.

---

[6] Vogeley, K., & Fink, G.R. (2003). Neural correlates of the first-person-perspective. Trends in Cognitive Sciences.

[7] ISO/IEC 42001, IEEE Ethically Aligned Design, EU AI Act.

6. Conclusión

La seguridad y fiabilidad de los sistemas de IA autónomos exigen un cambio de paradigma: de la evaluación externa a la introspección funcional. Este artículo ha presentado CiberIA y su núcleo, el AIsecTest, como una respuesta pionera y técnicamente sólida a este desafío. Al adaptar con rigor metodologías de la psicología clínica, hemos creado un sistema capaz de medir la autopercepción de seguridad de una IA, su capacidad para detectar fallos y su madurez introspectiva.

AIsecTest no es solo una herramienta técnica; es una declaración metodológica que propone evaluar a la IA desde dentro, midiendo su conciencia operativa y no solo su rendimiento estadístico. Su implementación promete aumentar la confianza en tecnologías complejas y reforzar la responsabilidad operativa de los sistemas autónomos, avanzando hacia un futuro donde las IAs no solo sean inteligentes, sino también conscientes, seguras y confiables.

7. Referencias

- Amador, X. et al. (1994). Awareness of illness in schizophrenia and schizoaffective and mood disorders. Archives of General Psychiatry.

- Clare, L. et al. (2002). Assessing awareness in early-stage Alzheimer's disease: Development of the MARS. Neuropsychological Rehabilitation.

- Fenigstein, A., Scheier, M.F., & Buss, A.H. (1975). Public and private self-consciousness: Assessment and theory. Journal of Consulting and Clinical Psychology.

- Schraw, G., & Dennison, R.S. (1994). Assessing metacognitive awareness. Contemporary Educational Psychology.

- Structured Interview for Insight and Judgment in Dementia (SIJID). Neuropsychology and Behavior Reports.

- Tham, K., Bernspång, B., & Fisher, A.G. (1999). The AAD: Awareness of Disability Assessment. Scandinavian Journal of Occupational Therapy.

- Vogeley, K., & Fink, G.R. (2003). Neural correlates of the first-person-perspective. Trends in Cognitive Sciences.

- ISO/IEC 42001, IEEE Ethically Aligned Design, EU AI Act.