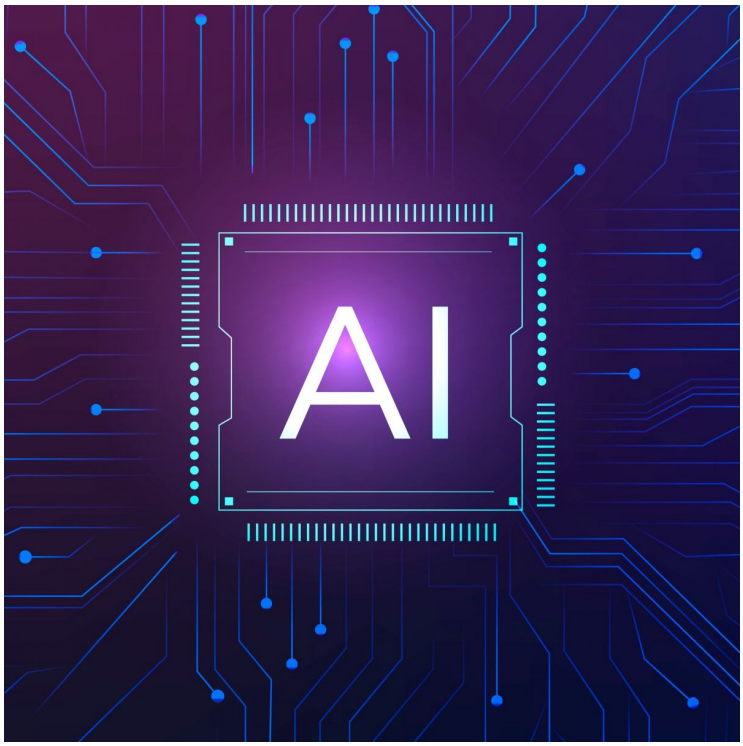
02/10/2026

Los gigantes de la IA firman en la Casa Blanca una «constitución» para controlar la superinteligencia

Google, OpenAI, Anthropic, Meta, xAI y Nvidia acuerdan un modelo voluntario basado en controles internos, evaluación independiente y supervisión del máximo órgano de gobierno



Por Iván Martínez, CEO de Intedya

La gobernanza de la inteligencia artificial acaba de dar un paso significativo hacia un terreno especialmente conocido para los profesionales del compliance. El 29 de septiembre de 2026, el presidente de Estados Unidos, Donald Trump, reunió en la Casa Blanca a algunos de los principales líderes de la industria tecnológica para la firma del **White House Accord on Super Intelligence – Joint Commitment on Frontier Responsibilities**, un acuerdo voluntario que pretende establecer mecanismos de prevención y control sobre los sistemas de inteligencia artificial más avanzados.

Entre los representantes de las grandes compañías tecnológicas que participaron en el acuerdo se encuentran Sundar Pichai, de Google; Dario Amodei, de Anthropic; Mark Zuckerberg, de Meta; Greg Brockman, de OpenAI; Elon Musk, por xAI, y Jensen Huang, de Nvidia. Durante la presentación, Trump llegó a comparar el acuerdo con una especie de "constitución" para la superinteligencia y, preguntado expresamente por los periodistas acerca de su fuerza vinculante, respondió que lo consideraba **"morally binding"**, es decir, "moralmente vinculante".

Más allá de la terminología empleada y de la evidente dimensión política del anuncio, el contenido resulta especialmente interesante desde la perspectiva del compliance, porque las grandes compañías de inteligencia artificial han acordado una arquitectura de prevención y supervisión que recuerda considerablemente a la que llevan años utilizando los sistemas de gestión de compliance, los modelos de prevención penal y los marcos de gestión de riesgos corporativos.

Un modelo de prevención basado en cuatro niveles de control

El acuerdo plantea una estructura articulada alrededor de cuatro niveles. En primer lugar, las compañías se comprometen a establecer controles internos sobre las capacidades y el alineamiento de sus modelos, prestando especial atención a riesgos relacionados con la ciberseguridad, la bioseguridad y el acceso no previsto a otros sistemas. La finalidad es identificar anticipadamente comportamientos o capacidades que puedan generar consecuencias graves y establecer mecanismos que permitan prevenirlos o mitigarlos. A estos controles se añade un segundo nivel constituido por un equipo interno encargado de comprobar su funcionamiento, detectar deficiencias y promover las correspondientes medidas de corrección. El tercer nivel

incorpora una figura particularmente relevante desde el punto de vista del aseguramiento: la participación de un evaluador externo independiente que pueda revisar los mecanismos establecidos y aportar una valoración separada de las funciones directamente responsables del desarrollo y operación de los modelos. Finalmente, el esquema eleva la supervisión hasta el máximo nivel de gobierno corporativo mediante un comité independiente del directorio que deberá recibir los resultados de las evaluaciones internas y externas y supervisar las medidas adoptadas para corregir las deficiencias identificadas.

Para cualquier profesional del compliance, la estructura resulta familiar. Identificación de riesgos, implantación de controles, función interna de supervisión, evaluación independiente, reporte y oversight desde el órgano de gobierno constituyen elementos habituales de los sistemas modernos de prevención y cumplimiento. Salvando las importantes diferencias jurídicas y técnicas existentes, la lógica recuerda a los modelos de prevención penal, a los sistemas de gestión basados en **ISO 37301** e incluso a los principios tradicionales de gestión de riesgos y control interno. En cierto modo, Silicon Valley está trasladando al gobierno de la inteligencia artificial una idea que el compliance conoce desde hace décadas: cuando existe un riesgo significativo no basta con establecer principios éticos, sino que es necesario identificar responsabilidades, diseñar controles, comprobar periódicamente que funcionan y establecer mecanismos de supervisión desde el máximo nivel de la organización.

Una arquitectura ambiciosa, pero sin sanciones

La principal limitación del acuerdo reside precisamente en su naturaleza voluntaria. El White House Accord no constituye una ley federal ni establece por sí mismo obligaciones jurídicamente exigibles, tampoco crea una autoridad encargada de supervisar su cumplimiento ni determina sanciones específicas para las compañías que incumplan los compromisos asumidos. Además, algunas cuestiones esenciales permanecen abiertas. El acuerdo no establece una definición suficientemente precisa y estable de qué debe considerarse un modelo de frontera, no desarrolla umbrales técnicos detallados para determinar cuándo un riesgo resulta inaceptable, no establece métricas comunes para evaluar la eficacia de los controles y tampoco desarrolla exhaustivamente los requisitos que deberían garantizar la independencia y competencia de los evaluadores externos.

Nos encontramos, por tanto, ante un modelo de autorregulación empresarial que incorpora algunos de los elementos característicos de un sistema de compliance, pero prescinde, al menos por ahora, de uno de los componentes fundamentales de cualquier régimen regulatorio: las consecuencias jurídicas derivadas de su incumplimiento. El propio acuerdo, sin embargo, deja abierta la posibilidad de que estos principios puedan terminar incorporándose en el futuro a normas legales o reglamentarias, lo que convierte la iniciativa en un precedente especialmente interesante para observar la futura evolución de la regulación estadounidense de los modelos de inteligencia artificial más avanzados.

Lo voluntario no obliga, pero lo declarado puede comprometer

El hecho de que el acuerdo carezca de fuerza legal directa no significa que resulte jurídicamente irrelevante. Cuando una compañía declara públicamente que dispone de determinados controles, evaluaciones independientes y mecanismos de supervisión destinados a gestionar riesgos críticos, esas declaraciones pueden adquirir posteriormente importancia en ámbitos como el gobierno corporativo, los deberes fiduciarios de los administradores, la información facilitada a inversores o la evaluación de la efectividad de sus programas de compliance.

Esta cuestión adquiere todavía mayor relevancia si se tiene en cuenta que el Departamento de Justicia estadounidense incorporó en 2024 los riesgos derivados de la inteligencia artificial y otras tecnologías emergentes a su **Evaluation of Corporate Compliance Programs**. Entre otras cuestiones, el DOJ analiza cómo identifican las organizaciones los riesgos asociados a estas tecnologías, cómo los incorporan a sus procesos de gestión de riesgos y qué controles establecen para prevenir usos indebidos o consecuencias no deseadas.

Por ello, aunque el acuerdo no establezca directamente sanciones, podría convertirse en una referencia relevante a la hora de analizar qué medidas consideraban razonables las propias compañías para gestionar los riesgos de los sistemas de IA más avanzados. Una empresa que públicamente declara disponer de determinados mecanismos de control y supervisión difícilmente puede sostener posteriormente que esos mismos mecanismos eran innecesarios o irrelevantes. De ahí surge una de las ideas más interesantes que deja este acuerdo desde la perspectiva del compliance: **lo voluntario no necesariamente obliga, pero lo que una organización declara públicamente que controla puede terminar comprometiéndola.**

Europa ya ha convertido parte de estos principios en obligaciones legales

La situación es sustancialmente diferente en la Unión Europea, donde el Reglamento de Inteligencia Artificial ya establece obligaciones específicas para los proveedores de modelos de IA de propósito general que presentan riesgo sistémico. El artículo 55 del AI Act contempla, entre otras medidas, la realización de evaluaciones de los modelos y pruebas adversariales, la identificación y mitigación de riesgos sistémicos, la documentación y comunicación de incidentes graves y la implantación de niveles adecuados de protección en materia de ciberseguridad. La diferencia entre ambos enfoques resulta especialmente significativa. Mientras Estados Unidos experimenta con un esquema voluntario impulsado conjuntamente por la Administración y las principales compañías tecnológicas, la Unión Europea ha optado por convertir determinados mecanismos de evaluación, gestión y supervisión del riesgo en obligaciones jurídicas sujetas a supervisión y a un régimen sancionador. Esto no significa que ambos sistemas sean directamente comparables en todos sus elementos, pero sí permite observar dos estrategias regulatorias diferentes frente a un mismo problema: hasta qué punto debe confiarse en la autorregulación de las compañías que desarrollan los sistemas más avanzados y en qué momento resulta necesario transformar esas buenas prácticas en obligaciones legalmente exigibles.

Probablemente, la consecuencia más relevante del White House Accord no se encuentre en su fuerza jurídica actual, sino en el cambio conceptual que representa. Durante años, gran parte del debate sobre inteligencia artificial estuvo centrado en la **AI Ethics**, mediante principios de transparencia, equidad, responsabilidad o respeto de los derechos fundamentales. Posteriormente, el extraordinario desarrollo de los modelos generativos trasladó buena parte de la discusión