48.ª Sesión Cerrada de la Asamblea Mundial de Privacidad (AMP)

Noviembre de 2026

Resolución sobre la IA agentiva que protege la privacidad

La presente resolución ha sido presentada por el Grupo de Trabajo sobre Ética y Protección de Datos en la Inteligencia Artificial (AIWG).

PATROCINADORES

- Comisaría de Protección de Datos Personales de Canadá
- Autoridad Catalana de Protección de Datos
- Oficina del Comisionado de Protección de Datos Personales de Hong Kong (China)
- Comisión de Protección de la Información Personal de Corea
- Agencia Española de Protección de Datos
- Comisionado Federal de Protección de Datos e Información de Suiza
- Oficina del Comisionado de Información del Reino Unido

COPATROCINADORES

- Oficina del Comisario de Información y Protección de Datos Personales de Columbia Británica (Canadá)
- Oficina del Comisario de Información y Protección de Datos Personales de Ontario (Canadá)
- Supervisor Europeo de Protección de Datos (Europa)
- Comisión Nacional de Informática y Libertades (Francia)
- Comisionado Federal Alemán para la Protección de Datos y la Libertad de Información [BfDI]
- Comisionado de Protección de Datos de Renania-Palatinado
- Comisionado de Información de la Isla de Man

La 48.ª Asamblea Global de Privacidad de 2026:

Reconociendo el surgimiento y la creciente adopción de los agentes de inteligencia artificial (IA) y de los sistemas de IA agentiva;

Entendiendo, sobre la base de una definición de la OCDE, que el término "agente de IA" se refiere a una entidad de software que percibe su entorno y actúa sobre él con cierto grado de autonomía, utilizando herramientas según sea necesario para alcanzar objetivos específicos y adaptarse a cambios en las entradas de datos y en los contextos;

Entendiendo, también sobre la base de una definición de la OCDE, que el término "sistema de IA agentiva" se refiere a un sistema compuesto por uno o varios agentes de IA capaces de dividir tareas, colaborar cuando proceda y perseguir de manera autónoma objetivos complejos durante períodos prolongados, diseñado para operar en entornos físicos o virtuales abiertos o menos previsibles y funcionar con una supervisión humana mínima;

Reconociendo que el componente fundamental en el que se basan la mayoría de los sistemas de IA agentiva actuales es un modelo de lenguaje de gran tamaño (LLM por sus siglas en inglés), que los desafíos en materia de protección de datos inherentes a dichos modelos también se aplicarán a los sistemas de IA agentiva, y que, dado que esta Resolución se basa en tal reconocimiento, algunas consideraciones podrían evolucionar si otras arquitecturas de base para los sistemas de IA agentiva llegaran a generalizarse;

Reconociendo, asimismo, que las autoridades individuales de protección de datos pueden perfeccionar ulteriormente estos conceptos, por ejemplo, definiendo los sistemas de IA agentiva como sistemas que desempeñan una función de orquestación (en lugar de coordinación) para los agentes de IA;

Limitando el alcance de la presente Resolución a aquellos sistemas y situaciones en los que un sistema de IA agentiva trate datos personales de una o varias personas, incluidos los casos en que dichos sistemas se utilicen para adoptar decisiones automatizadas sobre personas o para llevar a cabo acciones que las afecten;

Observando, asimismo, que, aunque cada vez se introducen más herramientas que permiten a las personas crear agentes de IA o utilizar sistemas de IA agentiva a título personal, el alcance de la presente Resolución se limitará al uso, la implementación o la distribución de agentes de IA y sistemas de IA agentiva en organizaciones públicas y privadas y por parte de estas;

Consciente de que un enfoque de la IA centrado en las personas puede generar importantes beneficios económicos, sociales y de interés público, entre ellos el aumento de la prosperidad y la respuesta a desafíos mundiales urgentes;

Destacando la importancia de garantizar que el uso de los sistemas de IA respete los derechos humanos, incluido el derecho a la no discriminación, que sea equitativo y lícito, y que prevenga los sesgos;

Reconociendo que la adopción de sistemas de IA agentiva podría aumentar el potencial de efectos adversos significativos asociados a los procesos de toma automatizada de decisiones, en particular si dichos sistemas reducen la supervisión humana, incrementan el sesgo de automatización o limitan la posibilidad de que las personas afectadas puedan impugnar de manera efectiva las decisiones;

Poniendo de relieve que las organizaciones siguen siendo responsables de los datos personales tratados mediante sistemas de IA agentiva (incluso cuando dicho tratamiento haya sido iniciado de forma autónoma por agentes de IA o resulte de su utilización de herramientas, capacidades, otros agentes o servicios conectados) y que las organizaciones no deberían asumir que esos sistemas reconocerán o cumplirán las normas organizacionales, las expectativas contextuales o las limitaciones aplicables;

Reconociendo asimismo que, cuando existan salvaguardas adecuadas y demostrables, los sistemas de IA agentivos pueden utilizarse para reforzar las prácticas de privacidad de una organización, al hacer que los procesos sean más eficientes, coherentes y auditables; apoyar la ejecución de tareas fundamentales (como la tramitación de solicitudes de acceso a datos por parte de los interesados); o reducir la frecuencia con la que los datos sensibles se ponen a disposición de personas dentro de la organización, garantizando al mismo tiempo que las decisiones sensibles desde el punto de vista jurídico o contextual sigan estando sujetas al criterio humano;

Reconociendo la importancia de comprender las características distintivas y las diferencias asociadas a las nuevas tecnologías que puedan dar lugar a nuevos desafíos en materia de privacidad y protección de datos;

Reconociendo que, si bien los sistemas de IA agentivos plantean preocupaciones nuevas y específicas en materia de privacidad, por lo general también suscitan muchas de las mismas cuestiones asociadas a otras formas de IA y que, por consiguiente, las resoluciones y declaraciones existentes de la Asamblea Global de Privacidad sobre esta materia siguen siendo pertinentes, entre ellas la Declaración sobre la Ética y la Protección de Datos en el Sector de la Inteligencia Artificial (40.ª AMP), la Resolución sobre la Rendición de Cuentas Responsable en el Desarrollo y la Utilización de la Inteligencia Artificial (42.ª AMP), la Resolución sobre Inteligencia Artificial y Empleo, la Resolución sobre los Sistemas de Inteligencia Artificial Generativa (45.ª AMP), la Resolución sobre la Recopilación, Utilización y Divulgación de Datos Personales para el Preentrenamiento, Entrenamiento y Perfeccionamiento de Modelos de IA, y la Resolución sobre la Supervisión Humana Efectiva de las Decisiones Tomadas mediante Sistemas de IA (47.ª AMP);

Poniendo de relieve asimismo la labor pertinente realizada por otras redes de autoridades de protección de datos, incluida la Resolución sobre la Gobernanza Ética de la Inteligencia Artificial de la Asociación Francófona de Autoridades de Protección de los Datos Personales (AFAPDP), las declaraciones en el marco de la Mesa Redonda de las Autoridades de Protección de Datos del G7 sobre la IA generativa, la IA y los niños, y el papel de las autoridades de protección de datos en el fomento de una inteligencia artificial digna de confianza, así como la Declaración conjunta sobre la creación de marcos fiables de gobernanza de los datos para fomentar el desarrollo de una inteligencia artificial innovadora y protectora de la privacidad, suscrita por 20 autoridades de protección de datos;

Haciendo hincapié en que, además de las consideraciones relativas a la privacidad expuestas en los documentos mencionados anteriormente, los agentes de IA y los sistemas de IA agentivos pueden crear, agravar o poner de manifiesto desafíos relacionados con la privacidad, entre ellos los siguientes:

- **Uso dinámico de herramientas**: Los agentes de IA que integran un sistema de IA agentivo, así como las herramientas, capacidades, complementos y demás componentes que utilizan o invocan, pueden haber sido desarrollados, suministrados o controlados por diversas

organizaciones y operadores, cada uno de los cuales puede aplicar prácticas de privacidad, condiciones contractuales y salvaguardas técnicas distintas, cambiantes o incompatibles entre sí. Además, mientras que en un sistema tradicional de IA el usuario puede desconocer de antemano la lógica que se aplicará para llegar a una conclusión, en un sistema de IA agentivo el usuario también puede desconocer desde el inicio qué componentes serán invocados o utilizados y, por tanto, cómo circularán los datos personales entre las distintas partes. Esta consideración se ve agravada por la fragilidad de los enfoques empleados por los sistemas de IA agentivos (es decir, pequeños cambios en una instrucción (*prompt*) pueden tener repercusiones significativas en su comportamiento). Esto plantea desafíos para las organizaciones en relación con el mantenimiento de la responsabilidad sobre el tratamiento de los datos, la gestión de las transferencias transfronterizas de datos, la provisión de una transparencia suficiente sobre las prácticas relativas a los datos y la garantía de que las personas puedan ejercer efectivamente sus derechos en materia de privacidad. La utilización de múltiples herramientas también genera la posibilidad de que se produzcan errores en cascada, en los que un único resultado incorrecto producido por una herramienta puede afectar negativamente a la exactitud de todas las demás herramientas.

- **Memoria**: Los agentes de IA pueden mantener una "memoria de trabajo", que puede incluir información sobre tareas anteriores (por ejemplo, para ayudar al agente a identificar los pasos más eficientes o eficaces para completar tareas similares en el futuro). Si dicha memoria contiene datos personales, pueden surgir riesgos relacionados con la existencia de datos inexactos o excesivos, la conservación de los datos más allá del período necesario para el fin previsto o la manipulación del agente de una manera que afecte a la confidencialidad, integridad o disponibilidad de esos datos. Cuando un mismo agente de IA es utilizado para múltiples tareas realizadas por varios usuarios, aumenta además el riesgo de divulgación accidental de los datos personales almacenados.
- **Acceso a los datos y autoridad delegada**: Los agentes de IA pueden tener acceso a múltiples fuentes de datos, interfaces, cuentas o sistemas del entorno en el que operan. Estas fuentes pueden ser internas o externas, o ser accesibles mediante el uso de herramientas. Por ello, es necesario considerar si el agente accede únicamente a los datos y capacidades mínimos necesarios para realizar una tarea; si está tratando innecesariamente datos sensibles; si los datos personales puestos a disposición para una tarea podrían combinarse con otros datos o reutilizarse con fines distintos o incompatibles; y, especialmente en el caso de datos externos, si los datos disponibles son suficientemente completos, exactos y fiables para el fin perseguido. Existe asimismo el riesgo de que los agentes sean manipulados para divulgar indebidamente información o conceder acceso a fuentes de datos a partes no autorizadas, modificar registros, cambiar permisos, comunicarse con terceros no autorizados o llevar a cabo otras acciones importantes e involuntarias o inapropiadas. La probabilidad de que esto ocurra aumenta cuando los componentes son inseguros, han sido comprometidos, están obsoletos o disponen de permisos inadecuados. Por último, la autoridad delegada a un agente

puede dificultar la distinción entre las acciones iniciadas directamente por un usuario y aquellas ejecutadas de manera autónoma por un agente de IA.

- **Seguridad**: El uso de herramientas y el acceso a datos propios de los sistemas de IA agentivos amplían la superficie de ataque y, en consecuencia, introducen vulnerabilidades de seguridad o las agravan. Por ejemplo, un sistema de IA agentivo puede incorporar datos procedentes de una fuente externa o el resultado generado por una herramienta de terceros al contexto utilizado para la siguiente etapa de un proceso, lo que crea posibilidades de ataques de envenenamiento de datos y de inyección de instrucciones (*prompts*). Los sistemas de IA agentivos con acceso amplio a los datos personales en poder de una organización pueden resultar objetivos especialmente atractivos para este tipo de ataques. La capacidad de los sistemas de IA agentivos para ejecutar acciones, por ejemplo mediante el uso de herramientas, también incrementa los riesgos de seguridad, ya que un agente comprometido que disponga de privilegios dentro de una organización podría ampliar rápidamente el alcance y el impacto de una vulneración. Los sistemas de IA agentivos también podrían utilizarse para aumentar la velocidad o la eficacia de los ciberataques.
- **Supervisión y revisión humanas**: Una característica definitoria de los sistemas de IA agentivos es su capacidad para influir en su entorno o ejecutar acciones de forma semiautónoma o totalmente autónoma. Por ello, con el fin de obtener los mayores beneficios o ganancias de eficiencia percibidos, las organizaciones pueden verse tentadas a reducir el número de ocasiones en que un sistema de IA agentivo interrumpe su funcionamiento para solicitar intervención humana, por ejemplo mediante aprobaciones. Esto puede disminuir la capacidad de ejercer una supervisión humana efectiva sobre los sistemas de toma de decisiones basados en IA o de permitir que las personas soliciten efectivamente una revisión humana de una decisión automatizada. Asimismo, la naturaleza semiautónoma o totalmente autónoma de los agentes puede hacer que sea difícil de explicar el proceso mediante el cual un agente ha llegado a una conclusión o decisión, así como verificar la validez y fiabilidad de dicha conclusión o decisión.
- **Implementación no autorizada o sin gobernanza**: Al igual que ocurre con otras herramientas basadas en la IA, las personas dentro de las organizaciones pueden considerar que la adopción de sistemas de IA agentivos aumentará su productividad. La creciente accesibilidad de estas herramientas también puede incentivar a las personas a implementarlas o utilizarlas con fines organizacionales sin la aprobación o supervisión adecuadas. Sin embargo, a diferencia de sistemas de IA más tradicionales, en los que una persona debe cargar datos de forma deliberada, la configurabilidad de los sistemas de IA agentivos y su capacidad para conectarse a datos, herramientas y sistemas organizacionales pueden exponer grandes volúmenes de datos personales a agentes de IA con pocos controles o sin ningún control. Tal utilización puede eludir las políticas organizacionales, los inventarios de datos, los controles de acceso, las salvaguardas contractuales y los mecanismos de respuesta a incidentes, y puede exponer datos personales o sistemas internos a agentes, herramientas o servicios externos no autorizados.

Reconociendo que las organizaciones tienen obligaciones en materia de privacidad cuando utilizan o implementan sistemas de IA agentivos y que, para cumplir dichas obligaciones y mitigar los efectos negativos de las consideraciones descritas anteriormente, deben aplicar medidas basadas en el riesgo. Entre dichas medidas pueden figurar las siguientes:

- **Establecer mecanismos adecuados de revisión y supervisión** a lo largo de todo el ciclo de vida de un sistema de IA agentivo, desde su preparación e implementación hasta su utilización y eventual retirada. Estos mecanismos pueden incluir evaluaciones de impacto relativas a la protección de datos, con el fin de garantizar que los riesgos para la privacidad se identifiquen y mitiguen antes de la implementación; auditorías técnicas de los sistemas de IA agentivos durante la contratación o antes de su implementación; y evaluaciones periódicas de la validez, exactitud, fiabilidad e idoneidad de los sistemas de IA agentivos. Estas medidas deben revisarse periódicamente, así como tras cambios significativos, incidentes o la identificación de vulnerabilidades emergentes. Asimismo, estas medidas podrían estar respaldadas por agentes de IA especializados.
- **Determinar los límites adecuados para los usos de los sistemas de IA agentivos**, acordes con la naturaleza del tratamiento, la sensibilidad de los datos y el posible impacto en las personas. En algunos casos, esto podría incluir limitar las acciones realizadas o las decisiones tomadas por los sistemas de IA agentivos a aquellas que puedan revocarse o revertirse sin causar perjuicios a las personas o causando un perjuicio mínimo.
- **Establecer políticas que regulen el acceso de los agentes a los activos de datos, así como su uso, conservación, eliminación y divulgación.** Estas políticas podrían incluir mantener inventarios de los datos a los que un agente puede acceder; limitar las circunstancias en las que un agente puede acceder, utilizar o divulgar datos sensibles; definir de forma clara los fines para los cuales los sistemas de IA agentivos pueden acceder a datos personales, tratarlos o utilizarlos; minimizar la cantidad de datos personales a los que un agente puede acceder para esos fines; implementar salvaguardas destinadas a impedir el uso de dichos datos para fines distintos; seudonimizar los datos antes de que un agente de IA pueda acceder a ellos; garantizar la exactitud de los datos a los que los agentes de IA pueden acceder; implantar mecanismos adecuados de validación y control respecto de las herramientas a las que los agentes de IA pueden acceder y controlar la cantidad o naturaleza de los datos que pueden revelarse a dichas herramientas; o garantizar que los datos personales tratados por el agente de IA o por el sistema de IA agentivo no se conserven durante más tiempo del necesario. Asimismo, estas políticas podrían contemplar arquitecturas de implementación seguras acordes con la naturaleza y los riesgos del tratamiento, incluidas medidas adecuadas y proporcionales para aislar o contener los entornos operativos de los sistemas de IA agentivos, limitar las exposiciones innecesarias a la red y evitar que los fallos o las vulneraciones afecten a otros sistemas o datos. Por último, las organizaciones también deberían contar con medidas para evaluar y garantizar el cumplimiento y la eficacia de estas políticas.
- **Garantizar la incorporación de los controles necesarios** en los agentes de IA y los sistemas de IA agentivos. Estos controles podrían consistir en medidas de privacidad desde el diseño y

por defecto que faciliten el ejercicio de los derechos de protección de datos de las personas (como los derechos de acceso y rectificación); exigir que el agente o el sistema de IA agentivo mantenga una trazabilidad integral de extremo a extremo (por ejemplo, registros de las llamadas a herramientas, de los datos de entrada y de los datos de salida); dar preferencia a sistemas de IA agentivos cuyas acciones estén diseñadas para ser repetibles o fiables; compartimentar la memoria de los agentes por usuario o desactivar, cuando proceda, el almacenamiento en memoria; establecer puntos de decisión o de acción clave (incluidos, cuando sea necesario, puntos intermedios dentro de un proceso de decisión antes de que se ejecute una acción) que obliguen al agente a solicitar intervención humana o, de otro modo, establecer límites a la autonomía de los agentes de IA o de los sistemas de IA agentivos, evaluando periódicamente la eficacia de tales medidas; o permitir a los usuarios y operadores pausar o detener la ejecución y, cuando sea factible y razonablemente viable, revertir acciones o restaurar sistemas y datos.

- **Llevar a cabo pruebas y supervisión en tiempo real** de los sistemas de IA agentivos. Las pruebas permiten verificar que los sistemas de IA agentivos se comportan conforme a lo previsto antes de su implementación y pueden incluir pruebas funcionales, pruebas adversariales destinadas a identificar vulnerabilidades y una validación rigurosa de salvaguardas como los controles de acceso a los datos, los mecanismos que desencadenan la intervención humana y los mecanismos de reversibilidad. Tras la implementación, a medida que los agentes interactúan con entornos dinámicos y en tiempo real, incluidos otros agentes, deben existir mecanismos para detectar comportamientos inesperados o fallos e informar sobre ellos.
- **Mantener registros** de las decisiones, acciones, llamadas a herramientas, datos de entrada, datos de salida e intervenciones humanas pertinentes, en un grado suficiente para respaldar la supervisión, la investigación y la rendición de cuentas. Las medidas de registro y trazabilidad deben cumplir ellas mismas los principios aplicables en materia de privacidad y protección de datos, incluidos la minimización de datos, la limitación de la finalidad, los controles de acceso, las salvaguardas de seguridad y los períodos de conservación adecuados, teniendo presente que dichos registros pueden contener instrucciones (*prompts*) y datos personales sensibles relativos a usuarios o terceros.
- **Mantener una gobernanza adecuada sobre la implementación, configuración y utilización de sistemas de IA agentivos dentro de la organización.** Esto puede incluir identificar a las partes responsables, respectivamente, del diseño, la utilización y los resultados del sistema de IA agentivo; mantener un inventario actualizado de los sistemas autorizados y de sus fines, modelos, herramientas, capacidades, fuentes de datos, conexiones y permisos; identificar implementaciones no autorizadas o modificadas de manera sustancial; establecer procedimientos para detectar, investigar y responder a incidentes; y actualizar los registros pertinentes tras cambios significativos.
- **Establecer mecanismos adecuados y proporcionales para la impugnación, la revisión humana y los recursos efectivos.** Cuando un sistema de IA agentivo adopte, recomiende o

ejecute una decisión o acción que produzca efectos jurídicos, materiales o de importancia similar para una persona, las organizaciones deberían proporcionar medios accesibles para impugnar el resultado y obtener una revisión significativa por parte de un responsable humano de la toma de decisiones debidamente cualificado y autorizado.

- **Implementar programas de alfabetización y sensibilización en materia de IA basados en las distintas funciones** dirigidos tanto a los usuarios finales de sistemas de IA agentivos como al personal responsable de la gobernanza, la contratación y la implementación, la alta dirección y los responsables de protección de datos, garantizando que cada grupo disponga de los conocimientos y competencias necesarios para comprender las capacidades, limitaciones, riesgos, requisitos de gobernanza e implicaciones para la privacidad asociados con la implementación y el uso de agentes de IA, así como las responsabilidades derivadas de la utilización de sistemas de IA agentivos. Estos programas pueden incluir conocimientos específicos de la organización, como las expectativas o buenas prácticas relativas a la formulación de consultas o la definición de tareas asignadas a un agente, las limitaciones respecto de los tipos de tareas para las que puede utilizarse un agente o los requisitos de documentación que deben cumplirse cuando se recurre a un agente en un proceso de toma de decisiones. Dichos programas también pueden favorecer el ejercicio de una supervisión humana efectiva al permitir que las personas responsables de revisar, validar, cuestionar o intervenir respecto de los resultados y decisiones de los sistemas de IA agentivos lo hagan de manera eficaz.
- **Proporcionar notificaciones, transparencia y explicaciones significativos a las personas afectadas.** Las organizaciones deben garantizar que las personas sean informadas, de forma clara, accesible y oportuna, cuando interactúen con un agente de IA, reciban datos generados por la IA o se vean afectadas por una decisión o acción adoptada o respaldada de manera sustancial por un sistema de IA agentivo. La información proporcionada debe ser acorde con la naturaleza y el posible impacto del tratamiento, y debe permitir que las personas comprendan el papel desempeñado por el sistema, los datos personales utilizados, los principales factores o procesos que condujeron al resultado correspondiente y la forma en que puede determinarse la responsabilidad por las acciones del sistema.

Por consiguiente, la 48ª Asamblea Mundial de Privacidad resuelve:

1. Instar a las organizaciones a que tengan en cuenta los riesgos para la privacidad identificados en la presente Resolución al diseñar o utilizar sistemas de IA agentivos, y alentar a las organizaciones que desarrollan dichos sistemas a investigar e implementar medios innovadores para abordar los desafíos relacionados con la privacidad.

2. Continuar, a través del Grupo de Trabajo sobre Ética y Protección de Datos en la Inteligencia Artificial de la AMP, desarrollando e intercambiando conocimientos y buenas prácticas en relación con:

a. las consideraciones en materia de privacidad asociadas a los sistemas de IA agentivos y las formas de mitigarlas;
b. los casos en que las organizaciones han implementado sistemas de IA agentivos para mejorar sus prácticas de protección de datos;
c. las formas en que los principios vigentes de privacidad pueden aplicarse de manera eficaz y previsible a los sistemas de IA agentivos; y
d. las maneras en que las autoridades de protección de datos pueden facilitar el desarrollo y el uso responsables de la IA agentiva mediante un enfoque regulatorio favorable a la innovación, incluso a través de la creación y operación de entornos controlados de pruebas (*sandboxes* regulatorios).

3. Examinar la posibilidad de realizar trabajos adicionales sobre esta cuestión, incluidos, entre otros, la elaboración de orientaciones para las organizaciones que deseen implementar la IA agentiva y actividades de sensibilización dirigidas al público (incluidas orientaciones para las personas que estén considerando utilizar sistemas de IA agentivos).

4. Seguir haciendo hincapié en la importancia de incorporar la protección de la privacidad desde el diseño en las tecnologías nuevas y emergentes, en particular cuando dependan del tratamiento de datos personales.

FTC abstención: La Comisión Federal de Comercio de Estados Unidos se abstiene de esta resolución.