La AEPD realiza un análisis de la Inteligencia Artificial Agéntica desde la perspectiva de protección de datos

Introducción

La Agencia Española de Protección de Datos (en adelante, "**AEPD**" o "**la Agencia**") ha publicado las orientaciones "Inteligencia artificial agéntica desde la perspectiva de protección de datos" (en adelante, "**la Guía**"), un documento de referencia que aborda los desafíos que plantea la integración de sistemas de IA agéntica en tratamientos de datos personales.

La Guía va más allá del análisis convencional de cumplimiento del RGPD identificando **riesgos específicos derivados de la arquitectura de estos sistemas y de sus características diferenciales**, proponiendo un marco de medidas de control adaptado a esta nueva realidad tecnológica.

En el presente documento analizamos las cuestiones clave abordadas por la AEPD y sus implicaciones prácticas para las organizaciones.

¿Qué es la IA agéntica?

La IA agéntica es definida por la AEPD como ***"sistemas basados en IA con capacidad de actuar de forma autónoma para conseguir el cumplimiento de objetivos".*** Esta definición está alineada con el modelo recientemente publicado por la Autoridad de Desarrollo de Medios de Comunicación e Información del gobierno de Singapur, que identificaba sistemas de IA agéntica como *"sistemas que pueden planificar múltiples pasos para alcanzar objetivos específicos, utilizando agentes de IA".*

Para ello, los agentes de IA descomponen tareas complejas en subtareas y las ejecutan de forma planificada según una cadena de razonamiento. En su implementación pueden utilizar distintas herramientas, y son capaces de entender el contexto mediante el acceso a servicios internos o externos a la organización.

¿Cómo impacta la IA agéntica en los tratamientos de datos personales?

Como apuntábamos, la **autonomía** es una característica propia de la IA agéntica. Por lo tanto, una consecuencia directa de la integración de estos sistemas en tratamientos de datos personales existentes o en nuevos tratamientos es el **cambio de paradigma en la concepción del control sobre las acciones del sistema**.

Se deben reevaluar cuestiones como: aparición de nuevos intervinientes; extensión del tipo y categorías de datos tratados, así como del número interesados sujetos al tratamiento; menor nivel de transparencia de los tratamientos; retención de datos en más intervinientes y sistemas; aparición de nuevas finalidades; desarrollo de acciones automatizadas con efectos en los interesados; o nuevos impactos y riesgos para los derechos y libertades fundamentales.

Ante este nuevo paradigma, la AEPD es clara: la **materialización de los riesgos** identificados "*no es consustancial al uso de sistemas de AI agéntica el que estas las produzcan, sino del tipo y forma de*

*configuración del **sistema empleado**, y de cómo se implementen **medidas** en el marco del tratamiento*."

Por lo tanto, el objetivo que persigue la AEPD con la Guía es facilitar un esquema general – no aplicada a tratamientos o IA agénticas específicas – que responsables del tratamiento puedan tomar como referencia a la hora de **identificar riesgos, vulnerabilidades** y los **aspectos a tener en cuenta** a la hora de readaptar sus estructuras internas de gobernanza y gestión de riesgos al uso IA agéntica.

A continuación, recogemos las cuestiones principales analizadas por la AEPD.

II. Marco de Gobernanza de la información y políticas internas de las entidades.

La AEPD destaca la importancia de contar con un **sistema de gobernanza adaptado** que incluya los sistemas de IA agéntica y que cubra al completo su ciclo de vida como la medida *"más importante que se pueda adoptar en una organización*".

En este sentido, la AEPD establece que "*La gobernanza ha de ser única, lo que es importante es garantizar que los elementos de gobernanza que surgen del uso de IA agéntica en los tratamientos se puedan "mapear" sobre los ya existentes o, en caso contrario, crearlos*."

La AEPD recomienda utilizar el **principio de fallo seguro**, por el cual los tratamientos que incorporan IA agéntica deben ser diseñados anticipando errores, brechas, sesgos y otros efectos no deseados. Ello permite incorporar, desde las fases de diseño e ideación del tratamiento, **mecanismos capaces de reaccionar** ante la materialización de estos fallos y reducir su impacto. Al ser una tecnología en constante evolución y desarrollo, se recomienda, además, la **monitorización continua** de los sistemas y tratamientos que incorporen IA agéntica, en base a criterios objetivos, métricas de funcionamiento predefinidas y evidencias.

III. Complejidad de la estructura de los sistemas de IA agéntica

Una de las características propias de la IA agéntica es su capacidad de descomponer una tarea en subprocesos y desarrollar todas las acciones requeridas para producir el resultado final de forma autónoma. Estas **cadenas de razonamiento o *pipeline*** del agente pueden variar, desde un plan rígido codificado hasta modelos conversacionales donde las decisiones dependen de interacciones y modelos de razonamiento. En todo caso, a mayor complejidad de dichas cadenas, mayor es el riesgo de inestabilidad en los comportamientos del agente.

Un fallo frecuente son los **errores compuestos**, que hacen referencia al *"fenómeno en el que la precisión de un agente de IA disminuye a medida que una tarea requiere más pasos".* Cuando esto ocurre se corre el riesgo de generar líneas de razonamiento no alineadas con las políticas de la entidad, acceder a datos personales excesivos, reproducir sesgos originados en procesos anteriores, y, en definitiva, de perder el nivel general de calidad esperado en el resultado final.

La Guía destaca la importancia de diseñar los agentes de IA **controlando las cadenas de razonamiento** "*con relación al tipo de herramientas de acceso a información interna/externa que se pueden invocar, el número de accesos que pueden realizar, …* (la) *depuración de los argumentos de las funciones para limitar la cantidad y categoría de datos accedidos, y …*(el) *filtrado y análisis de la información a que están accediendo con relación al tratamiento".* De lo contrario, se corre el riesgo de

incumplir "*el principio de minimización, de exactitud y de limitación del tratamiento, además de exponer la seguridad de los datos".*

La complejidad inherente a las cadenas de razonamiento se sustenta sobre la **combinación de distintos componentes**, entre otros, LLMs, bases de datos o el acceso a múltiples interfaces de servicios.

A nivel jurídico, esta complejidad se traduce, en cadenas de contratos en los que resulta necesario armonizar las limitaciones de las licencias de cada componente del sistema de IA agéntica. Se trata de un **contexto contractual de "*naturaleza dinámica*"** que, durante el ciclo de vida del sistema, puede implicar cambios unilaterales en los contratos, incluyendo modificaciones en el propio objeto contractual. Ello da lugar a la necesidad de realizar revisiones iterativas de los contratos cada vez que se produzca una actualización de los términos y condiciones, así como de los aspectos técnicos de los propios servicios, a fin de determinar cómo cumplir el RGPD.

La AEPD propone el establecimiento de **catálogos o listas blancas de servicios** como una medida que permita agilizar la toma de decisiones y adaptación técnica de los sistemas. Sin embargo, consideramos que la extensión de esta medida al análisis jurídico de cada componente permitiría minimizar el impacto que los trabajos a realizar desde los equipos jurídicos de las organizaciones pueden tener en el desarrollo a nivel técnico de los sistemas.

IV. Implicaciones del factor humano en la integración de la IA agéntica

La llegada de la IA agéntica aumenta las posibilidades de **automatización de tareas** hasta ahora realizadas por el ser humano, en especial, aquellas que resultan repetitivas y en las que el valor aportado por el ser humano es limitado.

Con la integración de la IA agéntica se potencia el **desplazamiento de rol del ser humano de la ejecución a la supervisión.** En este contexto, es tentador adoptar estructuras en las que el personal supervisor se constituya como el único garante frente a la materialización de los eventuales riesgos y fallos de los agentes de IA. Sin embargo, no hay que caer en el error de considerar la supervisión humana como la única barrera frente a los riesgos derivados de la IA agéntica.

En línea con todo ello, el impacto del factor humano debe ser observado desde tres perspectivas:

- **Uso no supervisado (shadow agentic AI)**

  Está demostrado que la **disponibilidad y fácil acceso** en el mercado a esta tecnología, es un factor de riesgo para las organizaciones, ya que cuando las organizaciones no se adaptan ágilmente a las nuevas tecnologías disponibles, los empleados tienden a **incorporarlas de forma autónoma y no supervisada.**

  Esta tendencia da lugar a problemas que ya han tenido que ser afrontados por las organizaciones a raíz de fenómenos como como BYOD ("Bring Your Own Device") o BYOAI (Bring Your Own Artifical Intelligence). Sin embargo con la aparición del **BYOAgentic ("Build Your Own Agentic o construye tu propia IA agéntica)** los riesgos de contaminación en la operativa de la organización se multiplican exponencialmente.

La incorporación de sistemas que no se alinean con las políticas y metodologías de la organización, **generan y potencian riesgos que caen fuera del radar del responsable**. Si no se conocen, no se mitigan, y su impacto potencial es mucho mayor.

- **Sesgo de automatización.**

Las capacidades que despliega un agente de IA pueden llegar a deslumbrar al personal encargado de su operación y supervisión. El empleado que no es consciente de las limitaciones y riesgos que se esconden detrás de la IA agéntica tiende a "***aceptar las decisiones del sistema sin un análisis crítico suficiente,*** *y refuerza el criterio de autoridad atribuido a la tecnología, especialmente cuando ésta opera con un alto grado de autonomía".*

Frente a este riesgo, es responsabilidad de la organización asegurar que en el proceso de implementación de la IA agéntica el personal usuario del sistema, así como los supervisores, cuenten con las **capacidades y la cualificación** suficiente para ello.

- **Atribución de excesiva responsabilidad al personal supervisor**

La identificación de empleados que adopten la responsabilidad de supervisar los resultados ofrecidos por el sistema de IA agéntica es una medida obligada para contar un modelo de gobernanza adecuado. Sin embargo, la AEPD es clara al señalar que, "*no pueden reemplazar la diligencia obligada del responsable del tratamiento en el diseño de éste y de la selección de la IA agéntica utilizada como medio".*

Quiere decir con ello que **no** cabe **descargar la responsabilidad** respecto de los fallos del agente exclusivamente en el supervisor cuando este deriva de un **problema sistémico más amplio**, es decir, cuando tiene origen en un error de diseño del sistema de IA, o incluso del tratamiento o del marco de gobernanza en general.

V. Gestión de riesgos y el papel del DPO

La gestión de riesgos constituye, en esencia, un "*análisis crítico del futuro impacto del tratamiento*". Tradicionalmente, este análisis ha descansado en la identificación de escenarios prospectivos y en la ponderación de hipótesis de materialización más o menos probables, atendiendo tanto a elementos técnicos como organizativos. Sin embargo, cuando el tratamiento incorpora tecnologías complejas, la pretensión de prever exhaustivamente todos los impactos potenciales sobre los derechos y libertades de los interesados se enfrenta a una limitación estructural al ser imposible delimitar los posibles riesgos con certeza.

La Guía establece que la gestión del riesgo en tratamientos que utilicen IA agéntica debe ser *"un* ***proceso proactivo para gobernar las incertidumbres*** *que amenazan los derechos y libertades de los sujetos de datos en un tratamiento de datos personales: hay que identificar, evaluar y priorizar los riesgos, para después coordinar esfuerzos y tomar decisiones para evitar o minimizar su probabilidad o impacto".*

Para ello, la Guía identifica:

- Las principales **amenazas** asociadas a la implementación de tratamientos que incorporan IA agéntica con incidencia directa en la protección de datos personales y en el cumplimiento de las obligaciones normativas asociadas, excluyendo de su análisis aquellas que pudieran afectar a otros objetivos organizativos como la ciberseguridad corporativa, la eficiencia operativa o aspectos financieros.

- Un conjunto de **medidas de mitigación**, orientadas a las singularidades del sistema de IA agéntica. Dichas medidas persiguen como objetivos principales: (i) implementar el cumplimiento normativo en materia de protección de datos, (ii) reducir los impactos críticos que pudieran surgir en el marco de análisis de proporcionalidad —como la evaluación del interés legítimo, la compatibilidad de fines o la EIPD—, y (iii) mitigar los riesgos para los derechos y libertades de los interesados.

Para la gestión de los riesgos es necesaria la confluencia entre capacidades técnicas y conocimientos de cumplimiento normativo. Por lo tanto, la **alfabetización en materia de IA agéntica** debe constituirse como una medida transversal a implementar a lo largo de la cadena operativa del tratamiento. Desde los puestos directivos hasta los técnicos supervisores del sistema, incluidos los responsables TIC encargados del desarrollo y mantenimiento del agente de IA, deben ser conscientes de las vulnerabilidades y amenazas inherentes a estos sistemas, así como capaces de identificarlas, a fin de garantizar el cumplimiento efectivo de la normativa de protección de datos personales.

En este marco organizativo, la AEPD pone en valor la figura del **delegado de protección de datos** como perfil clave en la integración de una visión jurídica con la capacidad de comprender la tecnología utilizada, actuando como **eje de conocimiento e interlocutor entre responsables jurídicos y técnicos de la organización**.

VI. Conclusiones

La incorporación de la IA agéntica requiere que las entidades revisen de manera integral sus marcos organizativos, técnicos y jurídicos. La Guía publicada por la AEPD recomienda un enfoque proactivo que permita una adaptación eficaz a las particularidades y complejidad de estos sistemas.

Para ello, las organizaciones deben:

- ✓ **Definir e implantar un marco de gobernanza** único e integral para la IA agéntica, que permita mapear e integrar los elementos de gobernanza necesarios sobre las estructuras ya existentes en la organización o, en su defecto, establecerlos.

- ✓ **Implementar controles reforzados** de acceso a la información, minimización de datos y revisión iterativa de las cadenas contractuales asociadas a estos sistemas, en atención a la complejidad estructural de su arquitectura.

- ✓ **Revisar y adaptar el marco de gestión de riesgos** de la organización para integrar las amenazas asociadas a la IA agéntica, incluido el uso no supervisado de estas tecnologías (shadow agentic AI), y ajustar los análisis de riesgos y las evaluaciones de impacto a la complejidad e incertidumbre inherentes a esta tecnología.
- ✓ **Reforzar el papel del DPO** como figura central en la supervisión del cumplimiento normativo, garantizando su participación desde la fase de concepción de iniciativas y su función de interlocución entre los responsables jurídicos y técnicos de la organización.

Para más información puede contactar con:

Carlos Rodríguez Sau
Socio | Andersen Iberia
carlos.rodriguezsau@es.Andersen.com

Antonio Cano
Director | Andersen Iberia
antonio.cano@es.Andersen.com

Iomar Ruiz
Asociada | Andersen Iberia
iomar.ruiz@es.Andersen.com