INTELIGENCIA ARTIFICIAL AGÉNTICA
DESDE LA PERSPECTIVA DE
PROTECCIÓN DE DATOS

Febrero 2026

RESUMEN EJECUTIVO

Un agente de IA es un sistema de inteligencia artificial que utiliza modelos de lenguaje para cumplir un objetivo. Estas orientaciones son una introducción a las cuestiones de protección de datos que pueden surgir cuando responsables y encargados de tratamiento decidan utilizar sistemas de IA agéntica para implementar tratamientos de datos personales.

El objeto de este documento no es analizar el cumplimiento de un tratamiento concreto que emplea agentes de IA, sino como gestionar las peculiaridades que se incorporan en un tratamiento por el hecho de implementarse total o parcialmente con agentes.

Conocer esta tecnología es clave para adoptar decisiones informadas y basadas en evidencia sobre su implementación en tratamientos de datos personales. No basta el conocimiento como usuario: es necesario comprender sus fundamentos, alcances, límites y la manera en que se aplica. Tanto el rechazo irracional de la IA agéntica como su aceptación acrítica en el tratamiento de datos personales pueden resultar perjudiciales. En particular, hay que aprovechar de forma proactiva las oportunidades que ofrece esta tecnología para una mayor protección de datos desde el diseño y como herramienta PET por sí misma.

El texto se estructura realizando inicialmente una breve descripción de qué son los sistemas IA agénticos. A continuación, se analizarán las posibles vulnerabilidades de estos sistemas que afectan al cumplimiento de protección de datos, los aspectos de cumplimiento de la normativa de protección de datos y las amenazas específicas que pueden aprovechar las distintas vulnerabilidades. Finalmente, el documento enumera medidas que podría adoptar un responsable o encargado para garantizar el cumplimiento de la normativa de protección de datos y reducir o eliminar los impactos que la IA agéntica presenta en su despliegue en tratamientos con relación a los derechos y libertades de los sujetos de los datos. Estos análisis se centrarán en lo que es más distintivo en la IA agéntica como sistema en un tratamiento de datos personales, más allá de las vulnerabilidades, amenazas y medidas que son bien conocidas de las inteligencias artificiales generativas, o de otros elementos que componen estos sistemas.

Palabras clave: Internet y nuevas tecnologías, machine learning, aprendizaje automático, inteligencia artificial, protección de datos desde el diseño y por defecto, decisiones automatizadas.

INDICE

I. INTRODUCCIÓN

La automatización de tareas es el uso de tecnologías para ejecutar actividades repetitivas sin intervención humana constante. Este enfoque transforma eficientemente procesos que antes se realizaban completamente de forma manual, liberando tiempo para tareas de mayor valor. La utilización de sistemas de automatización se puede encontrar desde entornos industriales a entornos de oficina, incluyendo cualquier otro sector productivo o de servicios.

El desarrollo de los grandes modelos de lenguaje (LLMs[1] por sus siglas en inglés) cambia completamente el paradigma de la automatización haciendo surgir el concepto de IA agéntica como sistemas basados en IA con capacidad de actuar de forma autónoma para conseguir el cumplimiento de objetivos: los agentes de IA. La integración de modelos de lenguaje supone un salto cualitativo en la eficacia y complejidad de las tareas que pueden llevar a cabo, lo que abre un universo de posibilidades para la mejora de los procesos empresariales y en las Administraciones Públicas. A su vez, el uso de sistemas que implementan el paradigma de la AI agéntica (agentes de IA) trabajando colaborativamente para automatizar múltiples procesos conlleva un cambio en la propia concepción de la implementación de los procesos, flujos de trabajo o *workflows* de las entidades, así como del uso de la inteligencia artificial generativa (en adelante IAG) en el entorno laboral.

La capacidad que tienen los sistemas de IA agéntica para operar con autonomía, enriquecerse con la información del entorno digital y ejecutar tareas complejas introduce nuevos retos en muchos aspectos, entre ellos, en el ámbito laboral, la gestión y control de la organización, de resiliencia, de seguridad (*safety*) y ciberseguridad, aspectos éticos, la posibilidad de fraude, sobre la imagen corporativa, etc., además de los relacionados con la protección de datos personales. También, como sistemas de inteligencia artificial en sí mismos y por su tratamiento de datos, pueden surgir obligaciones que se deriven de normas generales, como del Reglamento de Inteligencia Artificial[2] o el Reglamento de Datos, o de normas específicas por ámbito de aplicación.

En este documento se va a realizar una introducción sobre las cuestiones de protección de datos que pueden surgir cuando responsables y encargados de tratamiento decidan utilizar sistemas de IA agéntica para implementar tratamientos de datos personales.

Este documento no va a abordar el empleo de agentes de IA en el ámbito doméstico (aunque pueden existir también implicaciones de cumplimiento normativo), ni aspectos sobre el desarrollo o evolución de modelos de lenguaje[3]. Tampoco se aborda la cuestión

[1] Aunque nos referiremos a lo largo del texto a LLMs, los pequeños modelos de lenguaje o SML han demostrado su eficacia en la implementación de varios casos de uso de agentes.

[2] Art.3.1 del Reglamento de Inteligencia Artificial: "«sistema de IA»: un sistema basado en una máquina que está diseñado para funcionar con distintos niveles de autonomía y que puede mostrar capacidad de adaptación tras el despliegue, y que, para objetivos explícitos o implícitos, infiere de la información de entrada que recibe la manera de generar resultados de salida, como predicciones, contenidos, recomendaciones o decisiones, que pueden influir en entornos físicos o virtuales.

[3] Aunque se empleen los datos de los servicios de agéntica para entrenamiento de inteligencias artificiales.

de agentes de IA de en una organización en los que no hay tratamientos de datos personales[4].

Los agentes de IA son medios, sistemas, que permiten implementar tratamientos de datos personales introduciendo una mayor automatización. Un mismo agente de IA puede ser utilizado para implementar operaciones en distintos tratamientos de datos personales. Por otro lado, un agente de IA puede ser una parte de las operaciones de un tratamiento que incluya, para implementar otras operaciones, el uso de otros sistemas o de operaciones realizadas por un operador humano.

El objeto de este documento no es analizar el cumplimiento de un tratamiento concreto que emplea agentes de IA, sino como gestionar las peculiaridades que se incorporan en un tratamiento por el hecho de implementarse total o parcialmente con agentes. Distintos tratamientos y distintos tipos de agentes implementados en dichos tratamientos podrían tener distintas implicaciones en protección de datos. En el análisis que se realiza en este documento se estudiarán estas implicaciones de forma genérica, teniendo en cuenta que no son inherentes, ni forman necesariamente parte de su naturaleza, a todos los agentes ni a todo uso de la IA agéntica.

El texto se estructura realizando inicialmente una breve descripción de qué son los sistemas IA agénticos. A continuación, se analizarán las posibles vulnerabilidades de estos sistemas que afectan al cumplimiento de protección de datos, los aspectos de cumplimiento de la normativa de protección de datos y las amenazas específicas que pueden aprovechar las distintas vulnerabilidades. Finalmente, el documento enumera medidas que podría adoptar un responsable o encargado para garantizar el cumplimiento de la normativa de protección de datos y reducir o eliminar los impactos que la IA agéntica presenta en su despliegue en tratamientos con relación a los derechos y libertades de los sujetos de los datos. Estos análisis se centrarán en lo que es más distintivo en la IA agéntica como sistema, más allá de las vulnerabilidades, amenazas y medidas que son bien conocidas en los elementos que la componen, como LLMs[5], bases de datos, comunicaciones, etc.

Todo ello se realizará dentro de las limitaciones que impone una nueva tecnología en continua evolución y cuyo análisis está todavía en desarrollo.

II. AGENTES DE IA

Los agentes digitales, basados en software tradicional y sistemas de control son previos a la aparición de la IA, sin embargo, sus funcionalidades eran limitadas en comparación con lo que se puede conseguir con la AI agéntica.

La IA agéntica supone mucho más que emplear LLMs. Comprender su funcionamiento resulta esencial para crear el clima de confianza, a través de la evidencia, de que se han

[4] Por ejemplo, en Park, T. (2024). Enhancing anomaly detection in financial markets with an LLM-based multi-agent framework se describe un framework de múltiples agentes de IA basado en LLMs para detectar e interpretar anomalías en datos financieros del mercado, con especial aplicación a datos del índice S&P 500. El sistema automatiza la validación de alertas de anomalías al coordinar agentes especializados (conversión de datos, análisis experto, comprobación cruzada y resumen) para mejorar eficiencia y reducir intervención humana en la vigilancia del mercado financiero.

[5] Por ejemplo, no se entran a valorar aspectos como el entrenamiento de los LLMs, que requiere un análisis diferenciado.

dispuesto las medidas y garantías adecuadas que permita a responsables y encargados de tratamiento obtener lo mejor de esta opción tecnológica.

En el texto se realizará una descripción sobre qué es un agente y el concepto de la IA agéntica teniendo en cuenta que las clasificaciones siempre tienen un carácter formal, y que la realidad tecnológica es una escala de grises que, en este caso, cruza conceptos como el de LLM, IAG y los nuevos desarrollos que se puedan producir en el futuro.

A. AGENTE DE IA

Un agente de IA es un sistema de inteligencia artificial que utiliza modelos de lenguaje[6] para cumplir un objetivo[7]. Un agente de IA actúa de manera adecuada según sus circunstancias y sus objetivos, es flexible ante entornos y metas cambiantes, aprende de la experiencia y toma decisiones apropiadas dadas sus limitaciones perceptivas y computacionales[8]. Para ello, descompone tareas complejas en subtareas, que se ejecutan de forma planificada creando una cadena de razonamiento, cada una de ellas implementada con distintas herramientas y que perciben el entorno mediante el acceso a servicios internos y externos.

Los agentes de IA se podrían definir por las siguientes características de forma general (dependiendo del agente y en mayor o menor grado):

- Autonomía: poder operar sin intervención humana constante.
- Percepción del entorno: procesan entradas en tiempo real mediante sensores, interfaces con aplicaciones (APIs), cámaras, etc., para interpretar contextos dinámicos. La interacción con el entorno permite evitar el problema del "corte estático de conocimiento"[9] de los LLMs.
- Acción: además de generar salidas de texto, código o multimedia, pueden ejecutar acciones externas, como envío de información, interacción con usuarios, ejecución de código, ejecución de contratos, control de dispositivos, etc.[10]
- Proactividad: anticipan necesidades o problemas en lugar de solo reaccionar, pudiendo iniciar acciones por sí mismos.
- Planificación y razonamiento: permiten planifican secuencias de acciones para cumplir metas específicas, evaluando alternativas y priorizando resultados óptimos.

[6] En general, grandes modelos de lenguaje (LLMs) como término ampliamente aceptado aunque podrían ser otro tipo de modelos de lenguaje, incluyendo modelos multimodales o MLLM.

[7] ISO/IEC DIS 22989 3.1.1 Agente: entidad automatizada que percibe el entorno y ejecuta acciones para alcanzar sus objetivos. Nota 1: Un agente de IA es un agente que maximiza la probabilidad de alcanzar sus objetivos exitosamente mediante el uso de técnicas del IA.

[8] Russell and Novig *Artificial Intelligence: A Modern Approach, 4th ed., p. 34*

[9] Fecha límite fija hasta la cual el modelo fue entrenado o ajusta, limitando su conocimiento a información disponible antes de ese punto. Para superarlo, los chat de IA comenzaron a implementar lo que serían los fundamentos de la futura IA agéntica: herramientas de búsqueda en Internet, RAG o memoria a corto plazo.

[10] Tanto para la percepción como para la acción existen protocolos estandarizados como el MCP (Model Context Protocol) que permite conexión cliente (agente)/ servidor (el servicio al que se conecta) o A2A (Agent to Agent) que permite la comunicación entre agentes.

- Memoria y Adaptabilidad[11]: Permiten definir el contexto, acumular experiencias, ajustar comportamientos a las reacciones del usuario y mejorar iterativamente mediante retroalimentación o autoevaluación con memoria a corto y largo plazo.

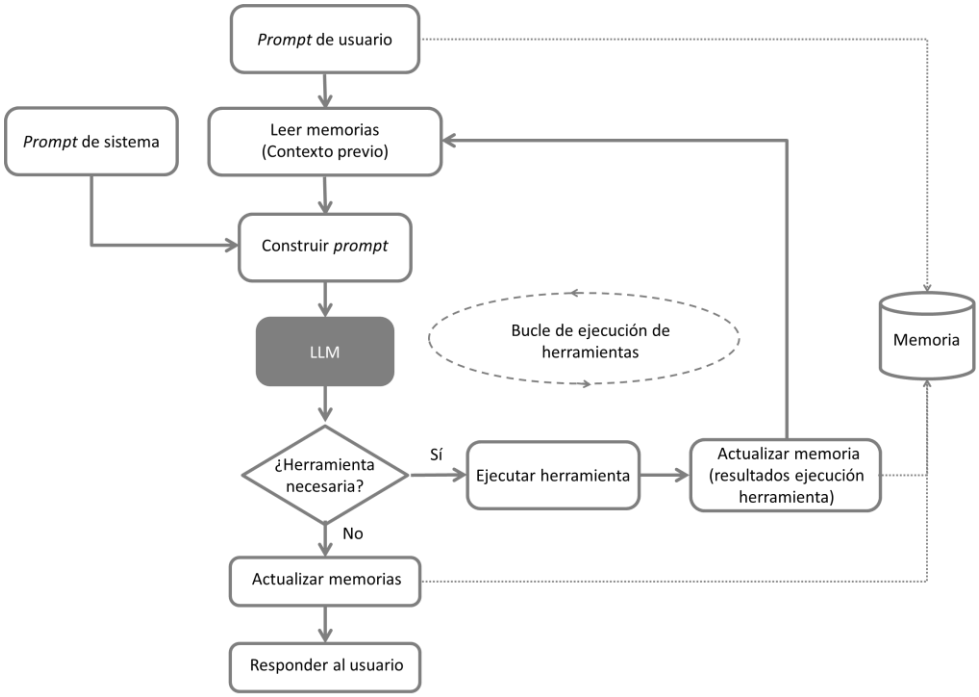

Figura 1 Ejemplo de una implementación básica de agente de IA

B. LA CADENA DE RAZONAMIENTO

La cadena de razonamiento, *pipeline* o procesamiento, es el proceso interno mediante el cual el agente descompone un problema en pasos lógicos sucesivos y encadenados, hasta llegar a una decisión o respuesta final. Esta cadena puede ser corta o, en los agentes que crecen en complejidad, muy larga (lo que se conoce con el nombre de *pipeline* largo) con múltiples etapas. Cada una de estas etapas puede involucrar distintos sistemas, formatos y niveles de confianza.

[11] En la literatura se habla de "aprendizaje" lo que puede llevar a confusión, en el sentido de que se esté reentrenando el/los LLMs que forman parte del agente. El "aprendizaje" del agente no se realiza por reentrenamiento del LLM. Aunque puede que se utilice la información para mejorar el LLM, no es una característica del agente, y en muchos casos no se realizará.

Figura 2 Ejemplo de cadena de razonamiento

La flexibilidad de la cadena de razonamiento puede variar desde un plan rígido codificado o máquinas de estado finito, hasta modelos conversacionales donde las decisiones dependen de interacciones y modelos de razonamiento.

En este último caso es cuando aparecen los LLMs como uno de los componentes nucleares de los agentes IA. En un agente de IA pueden aparecer distintos tipos de LLMs y AIG con distintos propósitos[12]: capacidades de conocimiento, generación de contenidos (como traductores, transcriptores, etc) y elaboración de razonamiento. Lo que resulta característicos de los agentes de IA es el utilizar LLMs como máquinas de razonamiento que dirigirán una actuación autónoma compleja, analizando las peticiones del usuario, respondiendo de forma secuencial a las entradas, procesando el resultado de distintos servicios y/o construyendo una respuesta final. Independientemente de utilizar en sistemas de IA agéntica servicios de LLMs como IAG

[12] Pero también pequeños modelos SML o grandes modelos de lenguaje multimodales MLLM

de contenidos o de repositorio de información, lo distintivo es utilizarlos para descomposición de tareas.

Conocer la cadena de razonamiento permitirá conocer el ciclo de vida del dato, la fuente del dato, la fecha y hora exacta de extracción, cuándo, dónde y por quién se produce su transformación, y cuándo, dónde, por quién y con qué finalidad y legitimidad se carga en un repositorio, se usa o se descarga desde un entorno a otro repositorio[13].

C. PATRONES DE LOS AGENTES DE IA

La arquitectura de los agentes, también llamados patrones (*patterns*), implementa un marco de razonamiento, es lo que permite planificar y ejecutar tareas complejas, combinando procesamiento del lenguaje natural, razonamiento simbólico, interacción con el entorno digital y planificación orientada a objetivos, lo que les confiere un grado de independencia operativa. Estos patrones podrán tener distintas configuraciones.

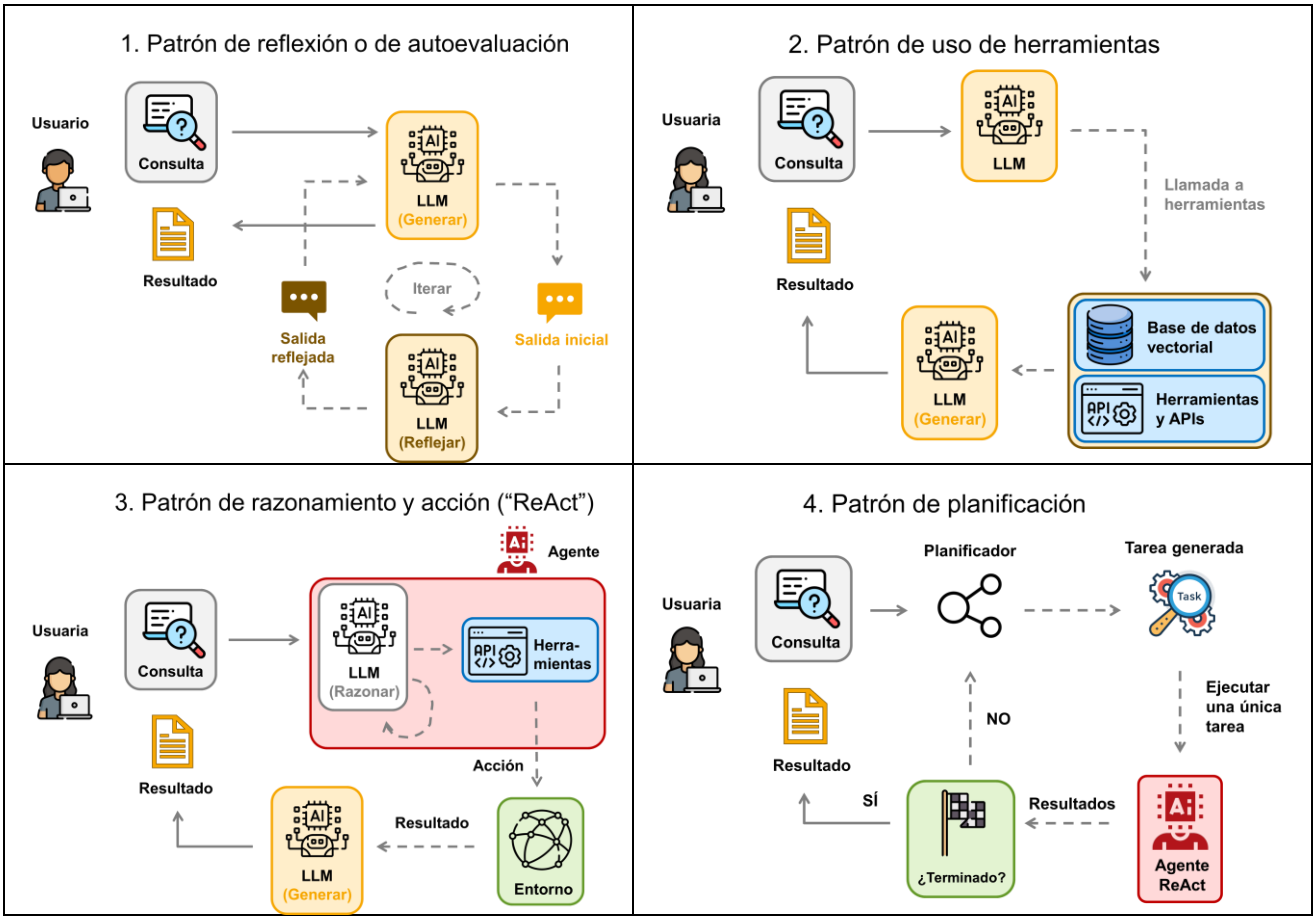

Figura 3 Representación simplificada de algunos tipos de patrones

[13] Este concepto es similar al que en logística diseña el flujo interno de materiales para el control de activos o el de gestión de flujo de producción. En el caso que nos ocupa, el dato, es un activo y un recurso.

De esta forma, los agentes de IA pueden automatizar tareas repetitivas de procesamiento de datos, analizar información para apoyar la toma de decisiones humanas o interactuar directamente con usuarios terceros y otros sistemas digitales.

A diferencia de los LLM, que son reactivos a acciones del usuario, los agentes pueden ser proactivos, y utilizar llamadas a/de herramientas en segundo plano para obtener información actualizada, iniciar operaciones, optimizar los flujos de trabajo y crear subtareas de manera autónoma con el propósito de alcanzar objetivos complejos.

Una de las características definitorias de los agentes de IA es su capacidad de realizar llamadas a servicios, es decir, conectarse a una API, a una base de datos, a sitios web o a otras herramientas, y utilizarla según sea necesario. Estos servicios pueden ser tanto remotos (por ejemplo, páginas o servicios web), como locales (por ejemplo, aplicaciones, capacidad de ejecución de código y datos almacenados en el sistema del usuario).

Aunque los agentes de inteligencia artificial operan de manera autónoma en sus procesos de toma de decisiones, dependen de metas y reglas previamente definidas por las personas. El comportamiento de un agente autónomo está determinado esencialmente por tres factores: el equipo de desarrolladores que diseña y ajustan el sistema de IA del agente; el equipo encargado de su despliegue y configuración; y, por último, el propio usuario, quien define los objetivos concretos que debe cumplir el agente y las herramientas a las que puede acceder para ello.

D. MULTIAGENTE

La arquitectura multiagente combina varios agentes, donde el comportamiento y las responsabilidades de cada uno están estrictamente definidos, comparten información y decisiones, y son capaces de colaborar, competir o negociar entre sí para alcanzar objetivos más elaborados.

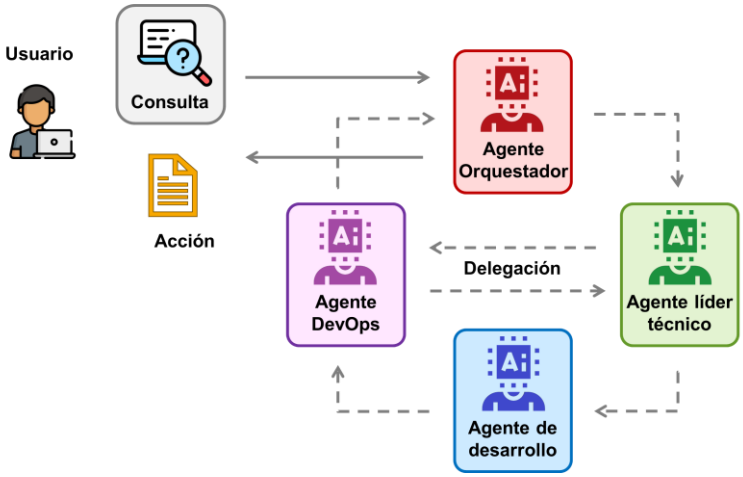

Figura 4 un ejemplo simplificado de arquitectura multiagente

Existen diversas aproximaciones a la IA multiagente: modelos centralizados, de ejecución secuencial de agentes, distribuidos o jerárquicos. Cada agente podrá tener un rango de acción (tareas que puede ejecutar y herramientas que puede invocar) y de autonomía distinto.

En el primer caso un agente central de planificación coordina el flujo de trabajo agéntico, mientras que los agentes operativos ejecutan sus porciones asignadas de la tarea, manteniendo su autonomía relativa. En cualquier caso, siempre se necesitará una capa de orquestación que coordine el ciclo de vida de los agentes, gestione dependencias, asigne roles a cada agente, establezca limitaciones por dominios y resuelva conflictos.

E. DETALLE DE LA ARQUITECTURA DE UN MULITAGENTE

La arquitectura general de un sistema de IA agéntica es un factor que se debe conocer para poder realizar un análisis de sus posibilidades, limitaciones y vulnerabilidades. Los componentes de un sistema de IA agéntica podría ser:

- Una aplicación que gestiona el interfaz para realizar tareas para el usuario o en nombre del usuario
- Uno o más agentes que implementarán distintos patrones de razonamiento y técnicas, como lógica basada en reglas, motores de flujos de trabajo deterministas, grafos de planificación, llamadas a funciones o encadenamiento de *prompts* que generalmente aceptan entradas en lenguaje natural, similares a las utilizadas por los modelos de PLN (Procesamiento del Lenguaje Natural). Estas entradas pueden ser *prompts* textuales y otros contenidos como archivos, imágenes, sonido o vídeo.
- Uno o más modelos LLMs (locales o remotos) se utilizan para el razonamiento, la generación de contenido final o intermedio, gestión de memoria e instrucciones para servicios.
- Los servicios, incluidos funciones integradas, herramientas locales y código de la aplicación, así como servicios locales o remotos.
- Los interfaces para el acceso e interconexión con herramientas y servicios externos (si es necesario): Internet, sensores, actuadores, etc.
- Almacenamiento externo para memoria persistente a largo plazo y memoria a corto plazo, incluyendo otras fuentes de datos, como bases de datos vectorial, repositorios de almacenamiento de objetos y contenido utilizado en *Retrieval Augmented Generation* (RAG).
- Servicios de soporte, que formen parte de la infraestructura del agente, como gestión de credenciales, control de accesos, trazabilidad de acciones, etc.

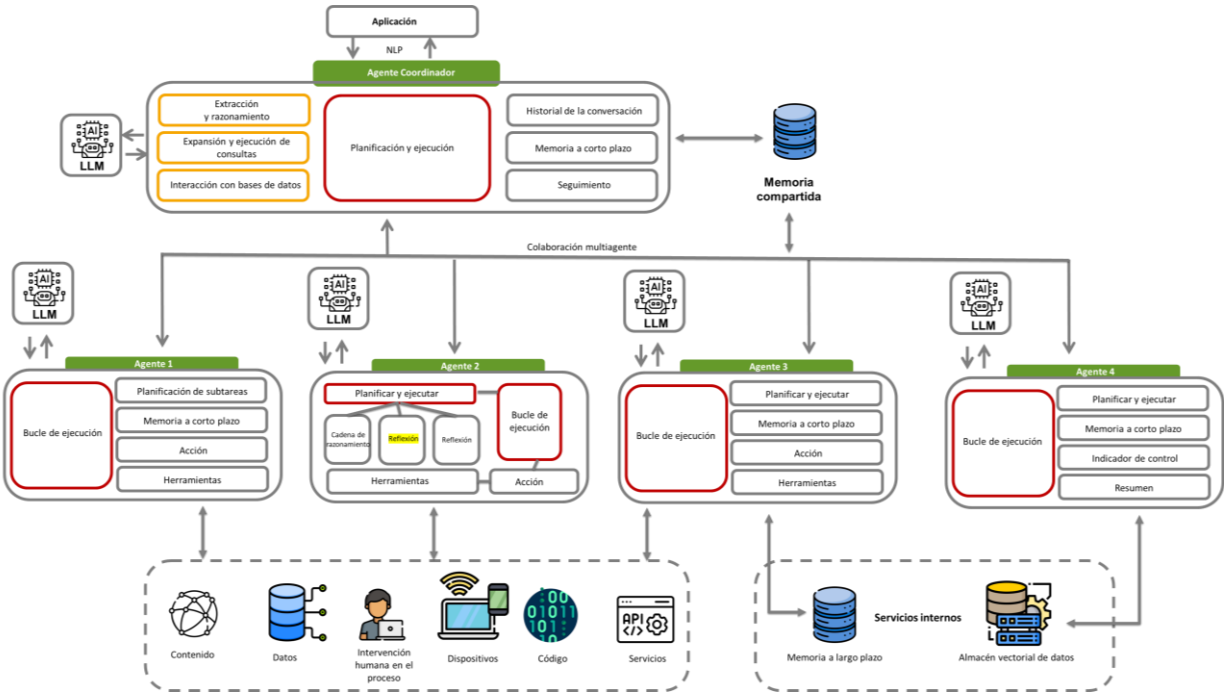

Figura 5 Detalle de la arquitectura de un sistema multi-agente de IA

Todos estos elementos se podrían implementar de forma local, sin acceso a servicios externos, o de forma totalmente externa, accediendo a un servicio de IA agéntica proporcionado por otra entidad. Entre estos dos extremos podríamos encontrar cualquier tipo de configuración que decida el responsable, con agentes cuya aplicación esté en local, pero donde parte de la memoria esté en la nube, con LLMs internos y servicios de LLM externos simultáneamente, etc.

III. AGENTES DE IA EN LOS TRATAMIENTOS

Los agentes de IA son medios que permiten implementar tratamientos de datos personales introduciendo una mayor automatización. Un mismo agente de IA puede ser utilizado para implementar operaciones en distintos tratamientos de datos personales. Por otro lado, un agente de IA puede ser una parte de las operaciones de un tratamiento que incluya, para implementar el resto de las operaciones, el uso de otros sistemas o de operaciones realizadas por un operador humano. Por ejemplo, en el caso de que sea necesaria la supervisión humana, en el tratamiento implementado con medios de IA agéntica tendrá que contemplarse dicha intervención desde el diseño.

Los agentes de IA son medios utilizados en un tratamiento que conforman su naturaleza, y también pueden alterar el contexto, el ámbito y añadir fines adicionales, además de alterar los riesgos inherentes al mismo. En el caso de tratamientos preexistentes, incluir IA agéntica obligará a una revisión de cumplimiento de dicho tratamiento. También se puede dar el caso que una entidad inicie nuevos tratamientos desde cero aprovechando las oportunidades que ofrece la IA agéntica, implementando con esta parte de los procesos del tratamiento.

Figura 6 Relación entre agentes y tratamientos

El objeto de este documento no es analizar el cumplimiento de un tratamiento concreto que emplea agentes de IA, sino sobre aspectos distintivo que podrían surgir con relación a protección de datos por el hecho de implementarse total o parcialmente con sistemas de IA agéntica.

A la hora de realizar dicho análisis es importante evitar la "niebla tecnológica" que puede provocar el solo nombrar IA agéntica a la hora de implementar un tratamiento. Por ejemplo, una agente puede implementar un proceso común en cualquier organización como sería organizar un viaje para un empleado. El agente, proactivamente apoyado en la IAG, cuando detectase en la agenda del empleado un viaje, desarrollaría un conjunto de tareas, como ponerse en contacto con diversos servicios de hospedería a través de Internet, comprobar el tipo de cambio de divisa, verificar el estado de las vías de transporte, gestionar los servicios para la adquisición de billetes transporte y obtener una previsión meteorológica actualizada. Teniendo en cuenta otros factores, (consultando las noticias), realizará una selección y se pondría de nuevo en contacto con los servicios, haría las compras oportunas y reenviaría la planificación y documentación al empleado.

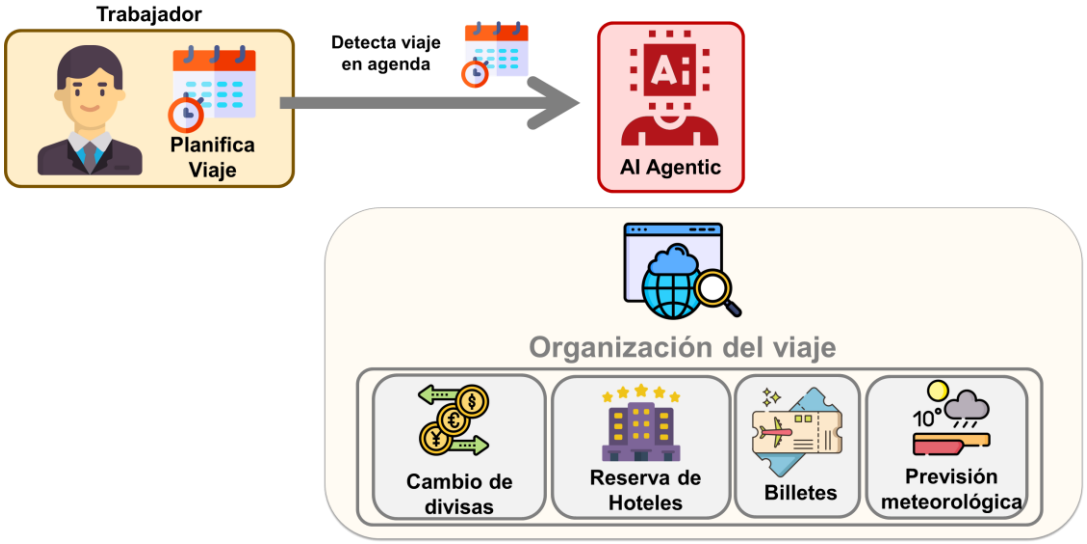

Figura 7 Ejemplo de gestión de viaje con IA agéntica

Tradicionalmente se ha utilizado los servicios de un administrativo que utilizaría los mismos datos y accedería a los mismos servicios, o bien la organización tendría contratada una agencia de viajes externa para realizar las mismas gestiones, incluso con la misma proactividad mediante acceso a la agenda del empleado. El análisis con relación a protección de datos (finalidad, minimización, legitimación, acceso a los servicios de Internet, etc.) será igual ya se implemente con un administrativo, o con la agencia de viajes contratada como encargado de tratamiento, o con un agente, incluyendo la capacidad o forma de determinar que hay un viaje en la agenda del empleado. Por ello, puede facilitar iniciar el análisis del cumplimiento el aterrizar en entidades o personas físicas las mismas operaciones que realice el agente y su relación con los servicios a los que accede (o la relación entidad externa que proporciona un agente-agencia de viajes[14]) para luego analizar los aspectos distintos que introduce la IA agéntica.

Con relación a esto último, la elección de uno u otro tipo de agentes para ser implementado en un tratamiento, podrían tener distintas implicaciones en protección de datos, incluso para distintos tratamientos[15]. En el análisis que se realiza en este documento se estudiaran estas implicaciones de forma genérica, teniendo en cuenta que no son inherentes, ni forman necesariamente parte de su naturaleza, a todos los agentes ni a todo uso de la IA agéntica.

Como se ha descrito anteriormente, la IA agéntica puede implicar la interacción con numerosos servicios internos y externos, a través de Internet, lo que expondría a los datos personales en una cadena de procesamiento en la que intervendría no solo el responsable, sino múltiples entidades bajo las políticas de privacidad, de cookies, de condiciones de servicio y contractuales de cada herramienta de terceros.

[14] Las garantías que ofrezca la agencia de viajes, ya sea atendida por personas o que realmente sea un agente, deberían ser las mismas.

[15] Un tratamiento podría ser, independientemente de los medios elegidos, de alto riesgo o tratar categorías especiales de datos, y otro tratamiento, que utilice como medio la misma IA agéntica, podría ser de bajo riesgo y no tratar ninguna categoría especial.

Por ello, a continuación se analizará:

- Qué nuevas vulnerabilidades, que afecten desde el punto de vista de protección de datos, podría implicar el incluir agentes de IA en tratamientos de datos personales.
- Qué aspectos de cumplimiento de protección de datos es necesario revisar cuando se plantean el uso de agentes de IA.
- Qué amenazas pueden explotar o materializar las vulnerabilidades detectadas con impacto en protección de datos.
- Qué medidas existen y están disponibles tanto para dar soporte al cumplimiento normativo, cómo para evitar los impactos críticos o gestionar el riesgo.

IV. VULNERABILIDADES Y TRATAMIENTOS DE DATOS PERSONALES

En este capítulo vamos a realizar un análisis preliminar de las vulnerabilidades más importantes que podrían surgir en un tratamiento por implementar operaciones con IA agéntica. Este análisis no es exhaustivo, entre otros factores porque está focalizado en aquellas que puedan tener impacto en el tratamiento de datos personales, y sean características del sistema agéntico como un todo, no de sus componentes individuales.

En la potencia y versatilidad de la IA agéntica reside, como en todo sistema complejo, sus principales vulnerabilidades.

Vulnerabilidad se define como la debilidad de un activo que pueda materializarse o ser explotada por una amenaza, potencialmente causando un impacto[16], en el caso que nos ocupa, con relación a la protección de datos personales.

Un sistema de inteligencia artificial agéntica integra diversos componentes de software, como son modelos de lenguaje[17], bases de datos, motores de planificación y otras herramientas analíticas. Asimismo, incluye interfaces tanto internas como externas que interactúan con múltiples servicios, los cuales, a su vez, pueden poseer sus propios niveles de conectividad. En consecuencia, todas las vulnerabilidades inherentes a cada uno de estos sistemas forman parte de las vulnerabilidades de la IA agéntica.

No obstante, resultaría inadecuado adoptar una perspectiva meramente aditiva, ya que la interacción entre los distintos componentes puede originar nuevas vulnerabilidades o amplificar las existentes, generando efectos multiplicativos. En definitiva, este tipo de sistemas introduce una superficie de ataque significativamente más amplia que la de los modelos de lenguaje, exponiendo al sistema a impactos y amenazas de mayor complejidad.

[16] ISO/IEC 27001:2022. Information security, cybersecurity and privacy protection — Information security management systems — Requirements

[17] Desde SMLs, LLMs, hasta MLLMs.

A. INTERACCIÓN CON EL ENTORNO

En el marco del tratamiento, la IA agéntica tiene la capacidad de interaccionar con el entorno para ejecutar todas o parte de las operaciones de tratamiento. La interacción puede limitarse a la propia organización, o se puede extender a servicios externos.

La invocación de herramientas y servicios en Internet son *de facto* salidas parciales que está realizando la IA agéntica al exterior. En particular, no están orientadas a la persona usuaria de la IA agéntica y no son el resultado final, por lo que podrían ser transparentes para dichos usuarios o para el responsable del tratamiento, pero contener datos personales o poder revelar información personal sobre las personas sujetas a dicho tratamiento.

- ***Acceso a datos de la organización y del usuario***

Con relación al apartado anterior, una de las funcionalidades comunes de la IA agéntica es el acceso a los servicios y datos internos con el propósito de enriquecer el contexto para la ejecución de tareas. Esta información podría ser la relativa a la persona usuaria, a un grupo de trabajo o de toda la organización. Algunos ejemplos podrían ser cuentas de correo electrónico, informes, decisiones, discusiones internas, reuniones, notas, conversaciones, una base de datos de clientes, etc. Esto supone el tratamiento de datos de los usuarios de la IA agéntica, que pueden ser datos personales de ese mismo usuario, como datos personales de otras personas, tanto de las personas cuyos datos son objeto de tratamientos, como de otras personas cuyos datos residen en los repositorios accedidos por la IA agéntica.

Un acceso no controlado y que no tenga en cuenta, no solo las políticas de compartimentación de datos de la entidad, sino las obligaciones de protección de datos desde el diseño y por defecto, podría producir tratamientos masivos de datos que irían en contra del principio de minimización, de limitación del tratamiento y de exactitud de la información sin son datos obsoletos o hay problemas de integridad. Si todo o parte de los componentes de la IA agéntica se implementan por encargados de tratamiento, podría suponer la comunicación de datos a terceros más allá de las finalidades del tratamiento. Además, en el acceso a conjuntos de datos no estructurados, parte de la información podría ser relevante, pero otra irrelevante o inadecuada desde distintos aspectos.

- ***Capacidad de percepción y acción externamente a la organización***

La interconexión a servicios de Internet permite la interacción de los agentes con el entorno exterior a la organización tanto para recabar información como enviar información (solicitudes, comandos o datos almacenados localmente) aumentando sus capacidades tanto de actuación, como de tratamiento de información.

La existencia de comunicaciones de datos bidireccionales con múltiples intervinientes, sin el necesario control de la entidad, puede incrementar en un grado notable las vulnerabilidades como el poder acceder por múltiples canales al control de la IA agéntica.

En cuanto a la información local que se envía al exterior, una libertad excesiva en la invocación de herramientas que recogen información interna podría provocar una comunicación de información innecesaria al no preparar el agente para discernir qué información es relevante de la que no lo es.

La conexión con el exterior no solo para ejecutar acciones, sino para obtener información podría estar utilizando fuentes inadecuadas, por falta de precisión, irreales, obsoletas, parciales, sesgadas o desinformadas. Sobre todo, si no hay procedimientos para la verificación de la fiabilidad, procedencia y coherencia de las fuentes utilizadas. De igual forma, si los comandos de solicitud de información no se han preparado adecuadamente, se puede estar recogiendo datos personales excesivos que no son relevantes para el tratamiento.

B. INTEGRACIÓN DE SERVICIOS

La IA agéntica se fundamenta en la integración de múltiples servicios. Formando parte de la IA agéntica, combinará el uso de al menos un modelo de lenguaje, herramientas de gestión de memoria y de ejecución de tareas. Externamente a la IA agéntica se deberá integrar con otros servicios como servidores de archivos, correo, servicios web, etc. Todos ellos podrán ser servicios locales o externos.

- ***Gestión de servicios***

Incluso cuando los servicios proceden del mismo proveedor, la naturaleza de la industria provoca muchas veces la evolución independiente de cada uno de ellos, con términos y contratos no homogéneos, incompatibilidades, discontinuidad de los servicios y cambios de interfaz. Esto implica una mayor complejidad en la gestión de herramientas, tanto para los servicios TIC de la organización, el usuario, y la gestión de respuestas por los propios agentes. También supone la creación de complejos flujos de datos y números sistemas que almacenan datos en reposo a corto y largo plazo (Ver apartado “IV.CC. Memoria”).

Todo ello puede implicar desafíos para el cumplimiento normativo de protección de datos, como por ejemplo la gestión de numerosos intervinientes, control de tratamientos adicionales, conservación de datos, ejercicio de derechos, problemas de exactitud, etc. También otros problemas funcionales como integración de APIs heterogéneas, latencias variables, colisiones de nombre, parámetros anidados, dependencia de esta mal interpretadas, confusión de los modelos al crear *prompts* que resultan confusos, de disponibilidad, de resiliencia, inestabilidades y falta de robustez, aparición de ciber-vulnerabilidades, incoherencia en los privilegios de acceso y uso, inestabilidad de la calidad del servicio, etc.

- ***Facilidad de desplegar servicios de IA Agéntica***

Existen servicios de IA agéntica que son fáciles de desplegar, intuitivos, con herramientas que permiten diseñar tareas y conectividad entre componentes de forma muy rápida incluso para usuarios finales. Este tipo de entornos son habituales en el

desarrollo de prototipos en software en otros contextos y facilitan el despliegue de sistemas como los agentes de IA.

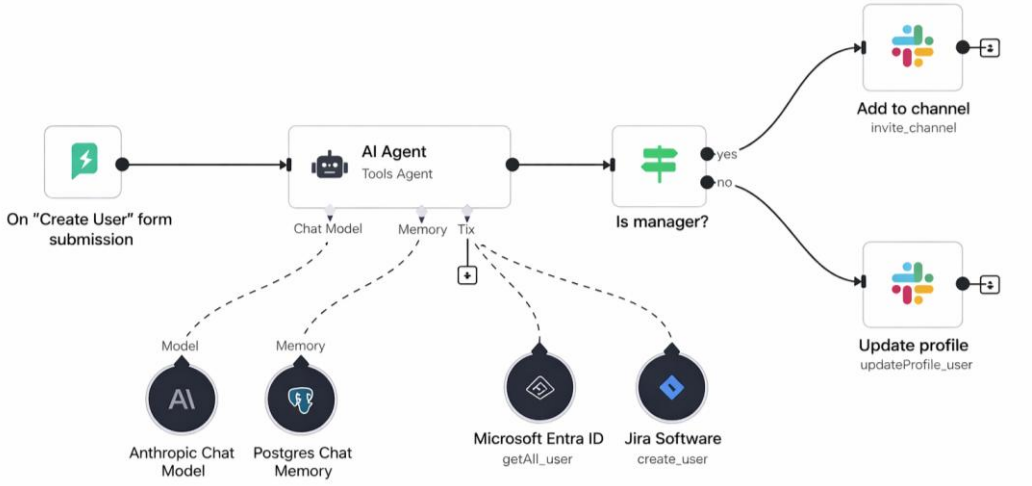

Figura 8 Entorno de desarrollo n8n (fuente: https://n8n.io)

Esto conlleva a la tentación de que usuarios no cualificados queden deslumbrados por sus posibilidades y realicen despliegues fuera de la gobernanza y las políticas de información de la entidad. La facilidad de tener una solución que parece que funciona con poco esfuerzo podría generar una idea de banalidad de las implicaciones e impactos para la protección de datos (y para la organización en general) al ocultar la complejidad inherente que tienen estos desarrollos en muchos aspectos.

Introducir un sistema de IA agéntica en el tratamiento de un responsable de tratamiento implica rediseñar un proceso de la organización en el que deberían intervenir, al menos, los responsables funcionales, TIC y de calidad, además del DPD cuando sea oportuno.

Los impactos de errores en el despliegue de sistema IA agéntica en un tratamiento pueden afectar desde la eficacia real al cumplimiento normativo, a la fiabilidad, la explicabilidad, la estabilidad y la robustez de los procesos, su escalabilidad y la disponibilidad, las vulnerabilidades que se generan en los tratamientos, el descontrol de los flujos de datos, la extensión y conservación de dichos datos, las consecuencias de brechas, la falta de preparación para la gestión de incidentes, etc.[18]

Si con los dispositivos móviles apareció en el ámbito laboral el problema del BYOD (*Bring Your Own Device* o traer tu propio dispositivo), y con los chat de IA el de BYOAI (*Bring Your Own Artifial Intelligence* o traer tu propia IA), con los IA agéntica aparece el problema de BYOAgentic (Build Your Own Agentic o construye tu propio flujo de proceso), al no contar con las políticas de la organización y los profesionales cualificados en gestión y técnicamente, y del uso de metodologías maduras en el diseño de procesos y aplicaciones.

[18] Una suerte de "AI slop" pero en la automatización de procesos.

C. MEMORIA

La memoria es una de las grandes ventajas de la IA agéntica y uno de sus elementos nucleares junto con los LLMs. La memoria en agentes de IA es la capacidad de almacenar y recordar contextos y experiencias pasadas para mejorar la toma de decisiones, la adaptación y el rendimiento. A diferencia de los sistemas que operan sin contexto, los agentes con memoria pueden reconocer patrones, adaptarse con el tiempo y utilizar retroalimentación previa, lo que resulta clave en aplicaciones orientadas a objetivos. Los modelos de lenguaje, por sí solos[19], no poseen memoria; esta debe integrarse como un componente adicional. Uno de los principales retos es gestionar la memoria de forma eficiente, almacenando solo la información relevante para mantener respuestas rápidas y un bajo nivel de latencia.

Hay dos tipos muy distintos de memoria en la IA agéntica. Una es la memoria que permite las funcionalidades de los agentes. La otra es la memoria que permite realizar la gestión de la IA agéntica y permite implementar mecanismos de control de la misma, y es la formada por todos los registros o logs de operación del sistema, de cada uno de los componentes del sistema, y los registros de los servicios accedidos por la IA agéntica.

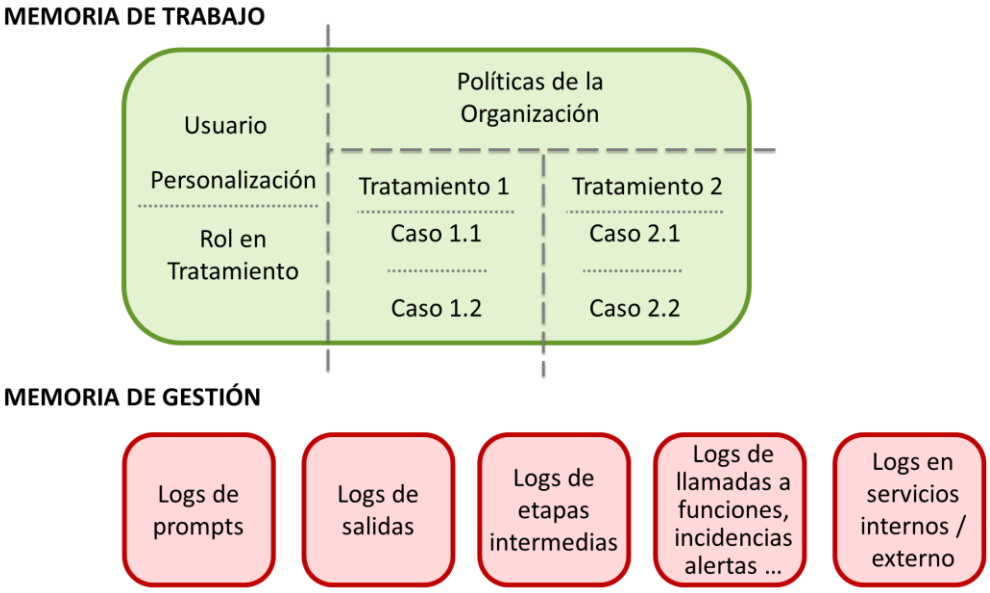

Figura 9 La memoria en la IA agéntica

- ***Memoria de trabajo***

Los agentes usan distintos tipos de memoria "de trabajo", a corto y a largo plazo (dependiendo del tipo de agente se podría hablar también de memoria a medio plazo). La aproximación más simplificada es que la memoria a corto plazo permite recordar interacciones previas dentro de un ciclo de ejecución[20] y la memoria a largo plazo permite

[19] Actualmente, los servicios que proporcionan acceso a LLMs sí incorporan memoria al evolucionar hacia el concepto de agentes, pero, formalmente, un transformer no tiene memoria tal como lo entendemos aquí.

[20] Por ejemplo, el historial de prompts del usuario en agentes conversacionales.

que los sistemas retengan información a lo largo de diferentes conversaciones o sesiones.

La memoria a largo plazo puede categorizarse a su vez como:

- Memoria semántica: implica la retención de hechos y conceptos específicos y se puede utilizar para personalizar aplicaciones recordando hechos de interacciones pasadas creando un "perfil" actualizado continuamente, con información específica sobre el usuario.
- La memoria episódica: permite recordar eventos o acciones pasadas y se utiliza para que el agente recuerde cómo realizar una tarea correctamente. Puede implementarse mediante *few-shot learning*, donde los agentes aprenden de secuencias pasadas que son usadas como ejemplos.
- La memoria procedimental implica recordar las reglas utilizadas para realizar tareas. Un enfoque eficaz para refinar estas instrucciones es la reflexión o *metaprompts* usando una IAG, donde el agente afina sus propias instrucciones basándose en sus interacciones.

La implementación concreta y las técnicas empleadas para mantener estas memorias puede ser muy diversa:

- Ficheros, bases de datos SQL o vectoriales.
- División en fragmentos y ventanas de contexto que pueda gestionar entradas complejas sin perderse, centrándose en las partes más relevantes.
- Incorporación de metadatos y etiquetado (fechado, usuarios, categorías, etc.) para filtrar rápidamente la información necesaria.
- Técnica de generación aumentada por recuperación (RAG) que permiten consultar un almacén de conocimiento en busca de contexto relevante antes de que el agente formule una respuesta.
- Técnicas de optimización de memoria: generación de resúmenes de información para ahorro de espacio, análisis y selección de información relevante, etc.

Desde el punto de vista de la implementación de un tratamiento en la organización usando sistemas de IA agéntica la información almacenada en la memoria se podría clasificar en:

- Memoria de la organización para todos los tratamientos: que es la información que la organización considera relevante para poder llevar a cabo la automatización en la organización. Esto contexto único podrá ser importante con relación a protección de datos para garantizar la coherencia y completitud (ver capítulo de Medidas).
- Memoria para cada tratamiento específico, que es relevante para un único tratamiento y no lo es para un tratamiento distinto. También puede contener información de contexto específica fijada por la organización.
- Memoria para cada caso tratado en el tratamiento, como podría ser un tratamiento que proporciona un trámite a un cliente y que no es relevante para

otro caso (dependiendo del tratamiento se podría dar una aproximación por caso o por cliente).

- Memoria de la persona usuaria para el mismo tratamiento, que pueden ser aspectos de personalización o también categorizada por tratamiento y por caso.

La organización lógica de la memoria podría tomar distintas formas, desde un gran repositorio en el que se vuelcan datos de las personas usuarias de la agéntica y de los datos personales de cada tratamiento, dejando a la agéntica el control de qué datos va a hacer uso en cada momento:

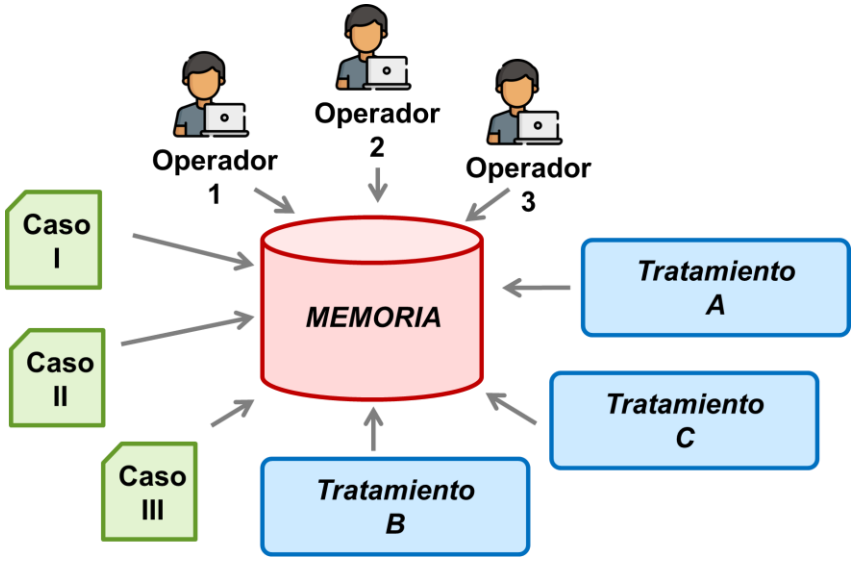

Figura 10 Organización de la memoria como un único repositorio lógico

En el otro extremo, la memoria puede dividirse lógica (o físicamente) para cada tratamiento, a su vez para cada caso y para cada persona usuaria que interviene en cada caso:

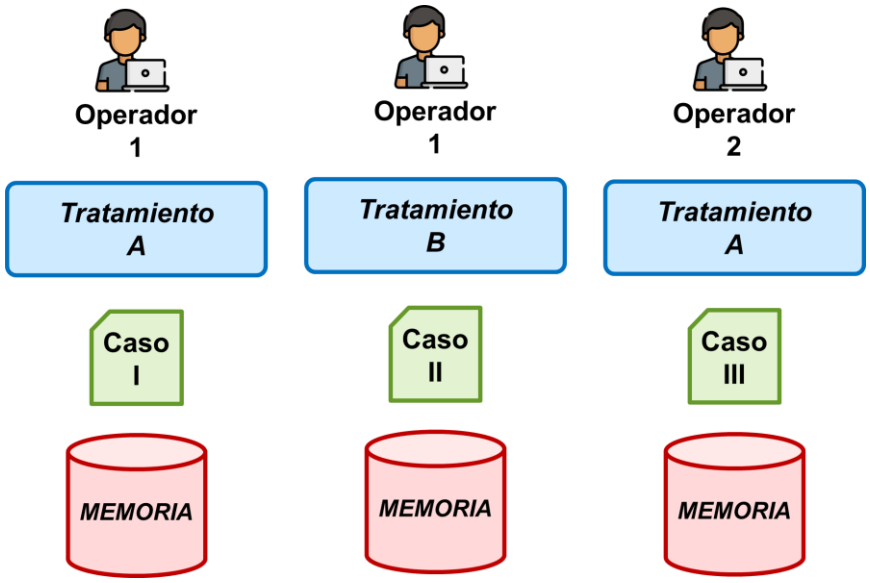

Figura 11 Distribución totalmente granular de la memoria

Entre estos dos extremos, se pueden encontrar varias soluciones de compromiso que se adecuen a las necesidades del responsable y de cada tratamiento.

La memoria, al igual que una gran ventaja, puede presentar vulnerabilidades con relación a protección de datos como:

- Relevancia: es necesario establecer políticas claras y eficaces de qué se ha de almacenar en la memoria para cada tratamiento. La relevancia puede estar descrita con *prompts*, cuando se analiza la misma con un LLM, o con otro tipo de técnicas. Entre otros, habría que garantizar que hay compartimentación entre contextos de distintos tratamientos (al menos), por ejemplo, que credenciales de usuario que se han proporcionado con un propósito no sean accesibles en el marco de otro propósito en el aparezca una petición de credencial[21].
- Coherencia y completitud del contexto: la información almacenada debe tener la calidad suficiente (incluyendo con relación a sesgos, con relevancia para dicho contexto, actualizada, sin contradicciones), en particular, si va a inferir o tomar decisiones sobre personas. Tanto en la memoria a largo plazo, como en los resúmenes elaborados en la memoria a corto plazo[22].
- Conservación: la información almacenada ha de ser la mínima necesaria para la operación del agente. No solo se trata de la información con relación a secretos comerciales o propiedad industrial, sino en particular de la información personal de la persona usuaria, de los sujetos sometidos a tratamiento o terceros, incluyendo información que puede inferir un perfil de cualquiera de ellos.
- Integridad: la información almacenada permite manipular el resultado de las inferencias y cambiar la actuación el propio agente, por lo que puede sufrir manipulaciones sobre el contexto y comandos o código de ataque que afecten a la confidencialidad, integridad o disponibilidad de los datos personales en poder de la organización (además de otros efectos que no son competencia de protección de datos).

▪ ***Memoria de gestión***

Hay que tener también en cuenta el impacto que puede tener la memoria en la sombra que es común en el uso de los sistemas digitales, como son los registros de actividad o log. Esta memoria también tiene su papel en la operación de la AI agéntica en cuanto tiene que ser explotada para analizar disfunciones, incidentes, ataques, alertas, etc.

La memoria almacenada en los registros puede ser, dependiendo de cómo se use, tanto una medida de privacidad como tener un impacto crítico o presentar un riesgo.

- Medida de protección de datos desde el diseño: al permitir auditabilidad de todas las acciones, trazabilidad, repetibilidad, exigencia de responsabilidad, medida disuasoria contra abusos, etc.

[21] Independientemente de la existencia de otros posibles controles.
[22] También se pueden estar aplicando en la memoria a medio y largo plazo.

- Impacto crítico, por ejemplo, cuando almacenen los registros información sobre las personas usuarias excesiva que se convierta en una hipervigilancia y vaya más allá de preservar privacidad y ciberseguridad.
- Riesgo en caso de que personas autorizadas a la gestión de dichos registros no cumplan sus deberes de confidencialidad, se produzcan tratamientos no autorizados por brechas de datos personales o por el uso de la información personal capturada para otras finalidades (por ejemplo, ajuste de LLMs).

Un aspecto singular aparece cuando un sistema de IA agéntica se utiliza para implementar distintos tratamientos. En ese caso, algunos de sus componentes, por ejemplo los LLMs, utilizarán registros o logs donde almacenarán la actividad de todos los tratamientos. Por ejemplo, almacenarán los prompts y las inferencias realizadas sobre todos ellos, convirtiéndose en nodos de recopilación de información personal de las personas cuyos datos son objeto de múltiples tratamientos, pero también podría almacenar datos de los

Esto podría tener un mayor impacto cuando dichos componentes son servicios externos a la infraestructura del responsable y gestionados por terceros, por ejemplo, cuando se utiliza el LLM somo servicio externo. En cualquier caso, se incrementa el riesgo de perfilado de los sujetos de los datos y el impacto en el caso de brechas de datos personales.

- ***Ejercicio de derechos***

En la medida en que la memoria del sistema de la IA agéntica almacena datos personales en el marco de uno o más tratamientos, y también registra qué accesos u operaciones se están haciendo sobre dichos datos personales, deberá contemplar desde el diseño la capacidad de ejercer todos los derechos del RGPD, entre ellos, acceso, rectificación, supresión, limitación y oposición.

D. AUTONOMÍA

La automatización agéntica implica que los agentes pueden actuar de manera autónoma, sin recibir instrucciones explícitas de un usuario humano. Esta autonomía permite decidir cómo se va a ejecutar la tarea, en qué pasos se subdivide, que fuentes internas o externas consultar, qué información tener en cuenta y cómo se tendrá en cuenta, tomar decisiones, ejecutar herramientas, realizar inferencias o generar resultados. Esta capacidad le da a la IA agéntica una gran capacidad de completar objetivos.

La autonomía es significativa en la actuación del agente de IA con el entorno: acceso y actualización de repositorios de datos, intercambio de datos entre intervinientes, combinación de dichos datos, gestión de otros procesos y generación de resultados, decisiones, evaluaciones u otro contenido generativo, todo ello internamente a la organización, como externamente a ella.

El nivel de autonomía del agente en el tratamiento es una decisión de diseño del responsable, y podría ser[23]:

- El agente propone, el humano opera.
- El agente y el humano colaboran.
- El agente opera, el humano es consultado o aprueba.
- El agente opera, el humano observa.

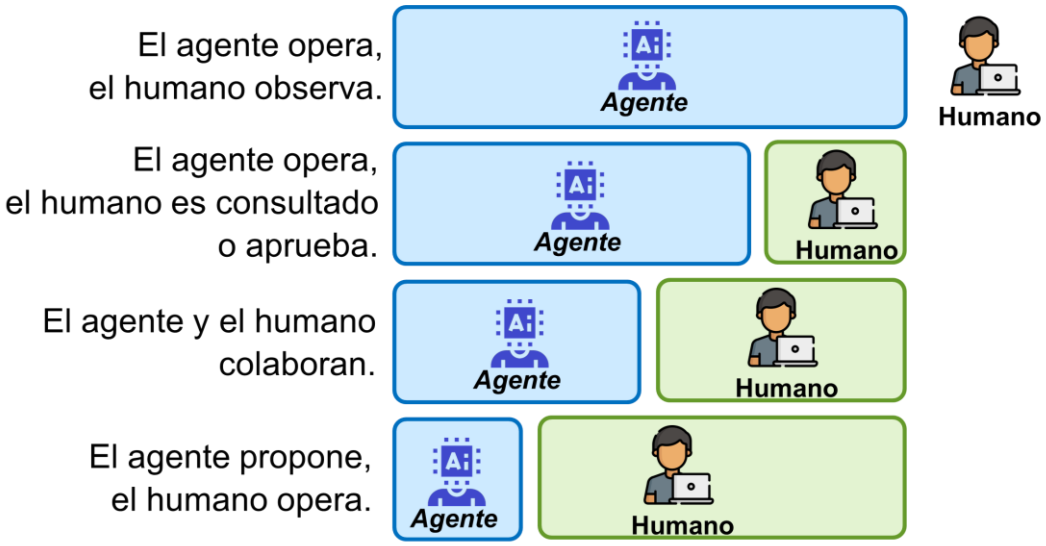

Figura 12 Niveles de autonomía de los agentes

Desde el punto de vista de la protección de datos personales hay varios aspectos que podrían ser afectados por dicha autonomía:

- Si los datos accedidos, actualizados o intercambiados cumplen los principios de minimización, exactitud, y limitación del tratamiento.
- Si dichas acciones son decisiones automatizadas de acuerdo con el artículo 22 del RGPD.
- Si dichas acciones tienen un impacto serio sobre el individuo y si son reversibles en el marco del tratamiento (acciones como borrar los datos de la persona física en los sistemas de la organización).
- Si existe la necesaria supervisión humana.
- Si realmente la tarea se ha organizado, subdivido y ejecutado de la forma correcta para garantizar que el tratamiento como un todo cumple realmente la finalidad.
- Si existe transparencia sobre su ejecución: calidad de los resultados, explicabilidad, repetibilidad, trazabilidad, auditabilidad y auditoría, entre otros.
- Si se proporcionan mecanismos de revocación en la IA agéntica y/o en el marco del tratamiento.

[23] En Feng et al. Levels of Autonomy for AI Agents (2025) https://arxiv.org/abs/2506.12469, se proponent cinco niveles.

- ***Transparencia y supervisión humana***

Los usuarios y desarrolladores pueden enfrentar dificultades para entender cómo algunos agentes de IA toman decisiones. Una falta de transparencia de los procesos de razonamiento internos (ya que las decisiones emergen de cadenas de inferencia distribuidas entre varios agentes y herramientas), y una limitada capacidad de comprensión por parte de los operadores humanos no suficientemente cualificados, puede generarse una confianza aparente, basada más en la percepción de correcto funcionamiento que en evidencias objetivas. Esta situación da lugar a una ilusión de fiabilidad, en la que el sistema parece consistente y eficaz, pese a no existir garantías sólidas sobre la validez de sus resultados.

El diseñar sistemas que constituyan cajas negras no es exclusivo de los agentes de IA, incluso se podría generar sin la inclusión de LLMs. No obstante, la velocidad y complejidad de los procesos de toma de decisiones de los agentes de IA pueden generar obstáculos más pronunciados para lograr una explicabilidad significativa y la transparencia necesaria para distintos objetivos: demostrar efectividad, evidencias de robustez, garantías para clientes, protección jurídica antes responsabilidades derivadas de las acciones, y, entre otros, cumplimiento de protección de datos en cuanto a derechos de los ciudadanos.

A su vez, se intensifica el sesgo de automatización[24], llevando a los usuarios 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.

Finalmente, la supervisión humana se vuelve más compleja, especialmente cuando no se han diseñado ni implementado mecanismos específicos que permitan, y en algunos casos obliguen, a una supervisión efectiva, continua y significativa.

- ***Planificación de tareas e interacción entre agentes***

Los mecanismos de descomposición de tareas y la interacción entre sistemas multiagente, junto con la orquestación de actividades entre dichos agentes, permite la ejecución de tareas de elevada complejidad añadiendo flexibilidad y adaptación que permite que la IA agéntica resuelva problemas en distintos contextos.

En la medida en que los agentes van a implementar tratamientos de datos personales en la organización hay que garantizar que todas las subtareas son las necesarias, solo las necesarias, y en el orden adecuado, teniendo en cuenta el impacto que pueda tener para los sujetos de los datos (y para otros objetivos de la organización). Existirán tratamientos en los que la descomposición y la orquestación será predefinida al menos hasta en determinado nivel. En otros, un LLM orientado a razonamiento puede hacer toda la descomposición.

[24] Elin Bahner, Anke-Dorothea Hüper, Dietrich Manzey, Misuse of automated decision aids: Complacency, automation bias and the impact of training experience, International Journal of Human-Computer Studies, Volume 66, Issue 9, 2008, Pages 688-699, ISSN 1071-5819, https://doi.org/10.1016/j.ijhcs.2008.06.001. (https://www.sciencedirect.com/science/article/pii/S1071581908000724)

Hay que tener en cuenta que un LLM no "razona", sino que extrae modelos de descomposición de tareas que se hayan incluido como datos de entrada en su proceso de entrenamiento[25]. Si se van a emplear para tareas muy complejas es importante garantizar que el LLM o SLM ha sido entrenado para ello. Por otro lado, que no hay posibilidad de contaminación entre distintos modelos aprendidos no compatibles (por ejemplo, subtareas de procedimientos administrativos de distintas jurisdicciones).

La complejidad técnica puede dar lugar a inestabilidad en el comportamiento emergente, es decir, a dinámicas impredecibles e indeseables propias de sistemas complejos, que no pueden anticiparse ni explicarse únicamente mediante el análisis de sus componentes individuales. Como consecuencia, pueden producirse resultados imprevistos o bucles de planificación infinitos.

Asimismo, la dependencia de llamadas secuenciales a herramientas externas puede generar bucles de ida y vuelta que acumulen latencia, especialmente en tareas de múltiples pasos en las que cada fase depende de los resultados de la anterior.

La indisponibilidad de un único proveedor, o la provisión de datos inconsistentes por parte de éste, puede desencadenar fallos en cascada que detengan operaciones críticas, comprometiendo la autonomía del sistema y la continuidad operativa.

Los errores compuestos (*compounding errors*) son el fenómeno en el que la precisión de un agente de IA disminuye a medida que una tarea requiere más pasos. Por ejemplo, un agente de IA consulta una base de datos sobre un sujeto con una *query* mal elaborada, recibe datos incompletos, los procesa como completos y realiza inferencias erróneas, que ocasionan la ejecución de tareas equivocadas, etc.

Con relación a los errores compuestos, tanto de las fuentes internas, de las fuentes externas o de resultados de inferencia intermedios, pueden generar información o líneas de razonamiento que no se ajusten a políticas de la entidad o exigencias regulatorias, como acceso a datos excesivos o de calidad insuficiente, inferencias erróneas, secretos comerciales, valores éticos de la organización, objetivos, sesgos, información financiera y, entre otros, consideraciones de protección de datos. Toda esta información puede producir resultados que se desvíen de la finalidad y creen perjuicios a usuarios, organizaciones, clientes o ciudadanos.

Una de las vulnerabilidades más críticas reside en la existencia de un único punto de compromiso (o *single point of compromise*, SPOC). Dado que estos sistemas están formados por agentes interdependientes, con un procedimiento de colaboración distribuida o planificación centralizada, que podrían comunicar a través de memoria compartida o protocolos de mensajería, la vulneración de un solo elemento de los antes citados puede comprometer todos los tratamientos que hacen uso de dicho sistema.

[25] Chengshuai Zhao et al. "Is Chain-of-Thought Reasoning of LLMs a Mirage? A Data Distribution Lens" 2026 https://arxiv.org/pdf/2508.01191

- ***Comportamiento no repetible***

La inferencia es el proceso mediante el cual los grandes modelos de lenguaje (LLM) y otros modelos de IAG, incluidos los sistemas basados en agentes, toman decisiones. En el aprendizaje automático clásico, la inferencia de tipo *one-shot* implica que un determinado dato de entrada produce una salida reproducible y determinista, siempre que se mantenga el control sobre el modelo y sobre las fuentes que alimentan los *prompts* de entrada.

En sistemas más complejos, como la inteligencia artificial agéntica, si no existe un control estricto sobre las fuentes de información, los servicios a los que se accede, sus versiones, la planificación de tareas, la memoria y los posibles comandos generados a partir de todos estos elementos, no es posible anticipar la salida del sistema.

Este problema no es inherente a la naturaleza de la inteligencia artificial agéntica, sino que responde a implementaciones concretas en las que no se dispone de un control adecuado del sistema construido. No obstante, mantener dicho control resulta más complejo debido a la mayor complejidad del propio sistema.

El agente puede generar cadenas de razonamiento y secuencias de acciones en las que pequeñas variaciones en un entorno no controlado (como por ejemplo un token diferente o un retraso en una llamada a una API) que desvían el plan completo, produciendo trayectorias distintas en cada ejecución. En este contexto, pueden aparecer bucles de razonamiento o bucles infinitos, cambios oportunistas de estrategia y comportamientos emergentes que no estaban explícitamente programados.

En consecuencia, la inferencia no predecible en sistemas de IA agéntica constituye un problema relevante, ya que impide anticipar y controlar con precisión el comportamiento del sistema cuando éste actúa en múltiples pasos, utiliza herramientas externas y se retroalimenta de sus propios resultados. Esta situación rompe muchas de las garantías clásicas de seguridad, responsabilidad, rendición de cuentas y cumplimiento normativo que se asumían en el software tradicionalmente controlado.

Asimismo, la falta de reproducibilidad de los errores dificulta la depuración en entornos que dependen de múltiples componentes, así como la implementación de políticas de seguridad para las personas (*safety*), de seguridad jurídica para las organizaciones y de protección frente a ataques (*security* y *cybersecurity*).

Finalmente, la verificación *ex ante,* incluidas auditorías, pruebas de regresión y procesos de validación regulatoria, se vuelve frágil, dado que un mismo dato de entrada puede generar, con el paso del tiempo, secuencias de acciones muy diferentes, lo que conlleva además una menor transparencia en los resultados finales.

- ***Capacidad de actuar en nombre del usuario o de la organización***

Muchos de los casos de uso de los sistemas de IA agéntica no podrían implementar automatización de forma efectiva si solo se pudiera interactuar con servicios, internos o externos, que no requieran autenticación. Por lo tanto, los agentes de IA deben ser capaces de solicitar y utilizar credenciales de usuario (por ejemplo, para acceder a su

correo o a su información en la nube) y credenciales técnicas o de máquina (por ejemplo, para acceder a cuentas corporativas en los LLM o servicios externos[26]).

La actuación en nombre del usuario no se limita al uso de credenciales. Una forma indirecta de dar privilegios que existe en algunas IA agénticas es incorporar herramientas que permiten al agente controlar el cursor y visualizar la pantalla del usuario, accediendo a los mismos datos y pudiendo realizar las mismas acciones que el usuario humano.

La concesión de permisos excesivos a los agentes constituye un factor crítico ya que un modelo con privilegios amplios puede ser utilizado como punto de pivote entre distintos sistemas, de modo que un compromiso inicial derive en accesos indebidos a bases de datos, servicios internos o credenciales sensibles, amplificando el impacto del incidente.

Asimismo, la ausencia de mecanismos de aislamiento entre servicios incrementa de forma significativa el riesgo de ejecución remota de comandos y código con niveles adecuados de privilegio a partir de entradas no confiables. Por ejemplo, el agente (a cambio de el acceso a servicios como son noticias, por error o motivado por un ataque) podría estar dando consentimientos o realizando contratos en nombre de una persona usuaria o estableciendo nuevas relaciones usuario-responsable.

Por último, la proliferación de identidades de máquina asociadas a agentes, servicios y automatizaciones genera un volumen elevado de cuentas técnicas cuya gestión y supervisión resulta compleja. Esta multiplicación dificulta la aplicación efectiva de controles de acceso, auditoría y revocación de privilegios, y aumenta tanto el riesgo de amenazas internas como la superficie de exposición frente a ataques externos, especialmente en entornos altamente distribuidos y dinámicos propios del agente de IA.

V. ASPECTOS DE CUMPLIMIENTOS DE LA NORMATIVA DE PROTECCIÓN DE DATOS

Un responsable de un tratamiento de datos personales podría elegir, entre los medios que va a utilizar para implementar el tratamiento, uno (o varios) sistemas de IA agéntica. Ya sea un servicio totalmente en local, totalmente en remoto, o cualquier fórmula intermedia, será un sistema que forme parte de la infraestructura tecnológica de una entidad, y podría implementar operaciones (incluso todas las operaciones) de uno o múltiples tratamientos.

Cuando los agentes de IA se utilizan en operaciones de tratamientos podrían surgir las siguientes cuestiones:

- Aparición de más intervienes que los responsable o encargados que forman parte original del tratamiento.
- Mayor extensión en el tipo y categorías de datos de los sujetos que eran objeto del tratamiento, incluyendo perfilados adicionales.

[26] Podría ser por acciones automatizadas de la propia organización, sin vinculación a una persona usuaria concreta, como respuestas automáticas, o bien porque internamente el agente registra un log de peticiones de usuario que mapea a accesos con una cuenta de la organización.

- Mayor extensión de sujetos de los que se tratan los datos, más allá de los sujetos que debían ser objeto de tratamiento, que podrían recogen del entorno al que puede acceder el agente o de la propia memoria del agente.
- Un mayor tratamiento de datos de las personas usuarias (empleados de la organización) que interaccionan con los agentes de IA.
- Menor transparencia en los tratamientos.
- Retención de datos en más intervinientes y sistemas.
- Nuevas finalidades.
- Acciones automatizadas con efecto en los sujetos de los datos.
- Nuevos impactos o riesgos para los derechos y libertades de los sujetos de los datos.
- Otros, dependiendo del tratamiento y de la forma de implementar la IA agéntica en el mismo.

La lista anterior no pretende establecer que cuando se utilice un sistema de IA agéntica en un tratamiento ocurran todas esas circunstancias. De hecho, 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.

A. DETERMINACIÓN DE RESPONSABILIDADES DE TRATAMIENTO

Responsable del tratamiento es quien, solo o junto con otros, determine los fines y medios del tratamiento, independientemente de la forma de que tengan dichos medios, ya sean sistemas de IA agéntica u otros. El responsable del tratamiento en el que se implementen sistemas de IA agéntica tendrá la obligación de:

- Garantizar el cumplimiento normativo.
- Gestionar los nuevos riesgos en el tratamiento que se podrían generar por el uso de sistemas de IA agéntica.
- Analizar la proporcionalidad de los impactos críticos[27] que podrían aparecer por usar agentes.

Cuando un agente de IA se ejecute de forma totalmente local no cabría más análisis de la responsabilidad. Sin embargo, en la mayor parte de los casos, los sistemas de IA agéntica accederán a servicios de terceros fuera de la organización para cumplir con su propósito. Estos servicios podrían ser los modelos de lenguaje, la gestión de la orquestación o incluso toda la IA agéntica como servicio proporcionado por otra entidad.

[27] Un impacto con certeza absoluta de ocurrencia se denomina impacto crítico. Por ejemplo, en un tratamiento de datos personales que es legítimo por alguna razón registrar todas la comunicaciones realizadas por una persona, tiene un impacto sobre sus derechos y libertades con certeza absoluta. Si no tuviera dicha legitimidad sería un incumplimiento normativo. Si se filtrasen los datos por una brecha habría un impacto adicional por lo cual hay un riesgo. Normalmente los impactos críticos se derivan de la propia definición del tratamiento y también se podría disminuir tomando medidas que lo hicieran proporcional a la finalidad del tratamiento, en muchos casos cambiando la definición del propio tratamiento, pero no es un riesgo, sino una certeza.

Los agentes de IA permiten la automatización de operaciones en el tratamiento con el soporte de IAG. Para estudiar las relaciones de responsabilidad de un tratamiento hay que analizar casuística que se encontrará el mismo responsable que se relaciona con otras entidades que prestarían el mismo servicio cuando implementa el tratamiento sin agentes de IA.

De esta forma, habría que analizarlo sin perjuicio de que hay tener en cuenta los tratamientos adicionales de datos que puedan producirse por uso de los componentes digitales que forman el sistema de IA agéntica. La complejidad de esta evaluación dependerá del nivel de automatización que se alcance dentro del tratamiento:

- Es posible que el agente acceda a servicios de terceros para obtener información no personal, como el horario de un servicio, el valor de activos financieros, datos históricos, etc. Además, la implementación del agente podría no permitir que dicho servicio los vincule a una persona usuaria concreta (no hay identificadores, ni cookies, ni histórico vinculado a la persona usuaria concreta ya que se filtra en el uso de la acción del agente). En dicho caso, la entidad que proporciona dicho servicio no tendrá ningún rol en el marco de protección de datos sin perjuicio de otros ámbitos normativos.
- Es posible que el agente envíe información no personal a servicios de terceros para realizar cualquier proceso: almacenamiento de información, traducción de textos, análisis de normas, elaborar un razonamiento sobre una tarea, etc. (donde podría intervenir un modelo de lenguaje). Si, como en el caso anterior, el servicio no lo vincula con una persona usuaria, no habría una relación de protección de datos. Sin embargo, si lo vincula con una persona usuaria con el objeto de prestar el tratamiento, por ejemplo, para guardar el contexto de las interacciones con el usuario, sería un encargado de tratamiento.
- En el caso anterior, si se da el caso de que sí se envíe información personal en el marco del tratamiento, dichos servicios actuarían como encargados de ese tratamiento[28].
- Es posible que el agente acceda a servicios de terceros para obtener información personal relativa a sujetos objeto de tratamiento u otros, por ejemplo, acceso a registros de administraciones públicas o entidades. Sin perjuicio de la legitimidad del acceso, la relación que se podría establecer sería una relación de comunicación de responsable a responsable. Sin embargo, si el agente está accediendo a los servicios de una agencia de coches de alquiler contratada por la entidad para obtener información de facturación de un empleado se trataría de una relación responsable-encargado.
- Es posible que el agente acceda a un servicio de terceros para transmitir información personal relativa a sujetos objeto de tratamiento. En ese caso hay que analizar la relación del responsable con la otra entidad independientemente del uso de los sistemas de IA agéntica. Por ejemplo, si un agente de un centro sanitario que presta servicios a una aseguradora respecto

[28] Párrafo 30 de las Directrices 07/2020 sobre los conceptos de «responsable del tratamiento» y «encargado del tratamiento» en el RGPD del Comité Europeo de Protección de Datos del 7 de julio de 2021

de los gastos de asistencia sanitaria que se hubieran llevado a cabo en el ámbito de un contrato de seguro, contacta con el servicio de la seguradora de forma automática para transmitir los datos, se trataría de una relación responsable a responsable[29].

- En el caso de que el propio agente sea un servicio proporcionado por otra entidad, y en la medida en que trate datos personales de las personas usuarias, y/o datos personales de clientes o ciudadanos en el marco del tratamiento original del responsable, la entidad que presta el servicio será encargada del tratamiento, como el caso del ejemplo anterior de agente-agencia de viajes.

En aplicación del principio de responsabilidad proactiva, el responsable deberá diseñar y documentar los flujos de datos del tratamiento, identificando para cada uno de los sistemas intervinientes los terceros implicados e identificando su rol dentro del marco de la normativa de protección de datos y el del resto de intervinientes.

En la medida en que la incorporación de los sistemas de IA agéntica implica la relación con servicios de Internet o servicios de otras entidades (ya sean encargados o terceros) aparecerá la misma casuística que surge cuando no hay automatización en el tratamiento:

- La utilización de datos personales proporcionados en el tratamiento para otros fines ajenos al responsable del tratamiento original. Por ejemplo, reentrenamiento de LLMs, seguridad, u otros.
- La creación de nuevas relaciones de responsabilidad con, dado el caso, las personas usuarias o aquellas personas cuyos datos personales se están tratando. Por ejemplo, a través de los propios interfaces de usuario solicitando su consentimiento para otros tratamientos.

En el primer caso, podría ser que esos tratamientos adicionales sean legítimos. El segundo caso también podría ser legítimo, en la medida en que no haya equívoco para el usuario de con quién está estableciendo la relación y cuando no se permita a la IA agéntica dar consentimientos de forma automatizada sin algún tipo de control.

En todos los casos, el responsable del tratamiento deberá tener la diligencia de controlar estas situaciones y esto dependerá del tipo de solución IA agéntica que se incorpore en el tratamiento. En el caso de que el sistema de IA agéntica sea implementada por el propio responsable se podrán tomar medidas para configurar los agentes de IA para poder determinar qué servicios se van a acceder (ver el capítulo “VII. Medidas”), evaluar los contratos o términos de servicio, revisar las cláusulas de protección de datos, cookies en su caso, deberá determinar la licitud de dichos tratamientos y las garantías de cumplimiento normativo, analizar el riesgo para los sujetos de los datos y evaluar si es proporcional el uso de dicho servicio o es más conveniente buscar alternativas. Es necesario evaluar el grado de cumplimiento normativo de las alternativas analizadas, en particular con relación, entre otros, al artículo 28 del RGPD, transferencias internacionales, conservación de datos, etc.

[29] https://www.aepd.es/preguntas-frecuentes/16-salud/1-salud/FAQ-1617-centros-sanitarios-y-hospitales-que-prestan-servicios-a-aseguradoras-y-mutuas-son-encargados-de-tratamientos-o-responsables

En el caso de que todo el servicio de IA agéntica sea encargado a otra entidad, se identifican las mismas obligaciones anteriormente citadas, además a las relativa a la cadena de subencargados.

Dependiendo del impacto del tratamiento en los derechos y libertades de los sujetos de los datos (y de otros intereses del responsable), será necesario recoger evidencias de cumplimiento más allá de los requisitos formales como, por ejemplo, realizando pruebas y estudiando incidentes que hayan podido reportar otros responsables.

En este punto cabe poner en valor la oportunidad que presenta la IA agéntica para garantizar el cumplimiento de contratos o términos de servicio de múltiples proveedores de forma proactiva y automática. En ese caso, la IA agéntica será una tecnología PET por sí misma, con aplicación en este campo como para cualquier organización que tiene que gestionar un entorno de múltiples servicios con actualización dinámica de condiciones legales.

B. TRANSPARENCIA

En el caso de que el uso de sistemas de IA agéntica en un tratamiento implique destinatarios adicionales de los datos a los previstos en el propio tratamiento, lo que ocurrirá en muchos casos, deberá informarse debidamente de su identidad. Si, por ejemplo, en el tratamiento esto supone que datos personales, ya sea de las personas usuarias o de las personas objeto de tratamiento, son remitidos a un servicio de IAG de otra entidad, se deberá informar adecuadamente a ambas categorías de personas.

Igualmente, deberá informarse de cualquier modificación que, debido al empleo de sistemas de IA agéntica en el tratamiento, de los plazos de conservación de los datos personales o, cuando no sea posible determinarlo con precisión, de los criterios utilizados para establecer dicho plazo. También de si se producen decisiones automatizadas adicionales (ver apartado “V.F. Automatización de las decisiones”) o transferencias internacionales adicionales (ver apartado “V.I. Transferencias internacionales”).

Cuando la incorporación en un tratamiento de soluciones basadas en agentes o sistemas de inteligencia artificial implique el tratamiento ulterior de datos personales recogidos con anterioridad para una finalidad distinta de aquella para la que fueron obtenidos, el responsable del tratamiento deberá informar al interesado con carácter previo a dicho tratamiento ulterior acerca de la nueva finalidad y de cualquier información adicional pertinente, de conformidad con lo dispuesto en el artículo 13, apartado 2, del RGPD.

Finalmente, la información deberá atenerse al objetivo de la información facilitada a los interesados tal y como se establece en el Considerando 39 del RGPD, conforme al cual las personas deben tener conocimiento de los riesgos, las normas, las garantías y los derechos relativos al tratamiento de sus datos personales, así como de los medios para ejercer dichos derechos.

C. LEGITIMACIÓN, MINIMIZACIÓN Y LEVANTAMIENTO DE PROHIBICIONES

La inclusión de sistemas de IA agéntica en un tratamiento podría implicar tratamientos adicionales de datos, aunque no necesariamente. Por ejemplo, si el administrativo que realizaba la gestión de viajes se reemplaza por un agente de IA en la propia organización, y esta tiene interfaces con los mismos servicios que se consultaban manualmente, el resultado será el mismo tratamiento de datos.

Es más, con el uso de la IA agéntica se puede obtener, o garantizar, un menor tratamiento de datos pues se podría dar el caso de que se han suprimido los tratamientos de datos de la persona usuaria cuando se accede a dichos servicios a través de Internet, ya que cookies o perfilados no podrían realizarse. En cualquier caso, el empleo de un agente de IA no constituye un fin en sí mismo.

Si la implementación de los sistemas de IA agéntica no implica tratamientos adicionales más allá del tratamiento original, no será necesario buscar legitimaciones para su inclusión en el tratamiento. Hay que tener en cuenta que, como los sistemas IA agéntica están compuestos de sistemas digitales siendo algunos muy complejos, incluirá más tratamientos de ciberseguridad que estarán amparados por el interés legítimo siempre y cuando sean orientados a dicho propósito, sean necesarios y proporcionales.

Si existen tratamientos adicionales, deberá tener su base legitimadora establecida y, en caso de categorías especiales de datos, una circunstancia que levante la prohibición. En el caso de que la base sea el interés legítimo, tendrá que superar la evaluación de que los fines de dicho interés legítimo estén claramente identificado, el tratamiento sea necesario para los fines del interés o los intereses legítimos perseguidos, y evaluar que el interés o los intereses legítimos no se vean anulados por los intereses o los derechos y libertades fundamentales de los interesados (también denominada «prueba de ponderación»).

En el caso de estar basado en el consentimiento, podría plantearse que sean necesarias medidas para la gestión de dicho consentimiento (ver apartado “VII.G. Gestión del consentimiento”).

La minimización de datos ha de contemplarse desde el diseño de los tratamientos y trasladarlo al diseño o configuración de los agentes. Por ejemplo, supongamos que en el marco de un tratamiento es necesario determinar si un empleado está en la lista de invitados de un evento. Para ello se podría descargar la lista de invitados y tratar los datos personales del empleado y, colateralmente, de todos los demás presentes en dicha lista. Cabría plantearse si es posible conseguir la misma finalidad sin tratar toda la lista de invitados (por ejemplo, preguntando al organizador si está el empleado en dicha lista) o incluso sin exponer de ninguna persona con estrategias de “conocimiento-cero”. En este ejemplo no se ha establecido si se está realizando el tratamiento con un operador humano o con una IA agéntica. En ambos casos, la minimización depende cómo se diseñe el tratamiento y qué instrucciones se han dado (en ambos casos) para el cumplimiento normativo.

La limitación del tratamiento también hay que abordarla desde el diseño. Por ejemplo, sea un tratamiento de atención al cliente en que la acción de la IA agéntica se

inicia ante una reclamación de una persona física, probablemente se va a realizar un tratamiento de datos personales. Parte o toda es interacción se podría almacenar en la memoria a largo plazo, pues podría ser necesario almacenar casuísticas concretas para dar una mejor respuesta en el futuro. Hay que plantearse si es necesario almacenar datos personales de pasados clientes en la memoria que va a utilizar la agéntica para realizar futuras acciones. En su caso, también hay que plantearse la legitimidad de tratar datos personales de otros clientes que pudieran quedar almacenados en la memoria a largo plazo en la elaboración de cada ejecución de la IA agéntica.

D. REGISTRO DE ACTIVIDADES DE TRATAMIENTO

El registro de actividades de tratamiento (RAT) es una herramienta fundamental para la gestión de cumplimiento de la normativa de protección de datos pues supone el catálogo de procesos de datos personales.

Cuando en un tratamiento se decida sustituir medios tradicionales por automatización agéntica deberá realizarse una actualización del RAT para determinar, por ejemplo, si esto supone una modificación en las categorías de datos personales objeto de tratamiento, si es necesario actualizar la información relativa a las categorías de destinatarios a los que se hayan comunicado o se prevea comunicar los datos personales, incluidos los destinatarios ubicados en terceros países u organizaciones internacionales, si se deben detallar nuevas transferencias de datos personales, modificar los plazos de conservación o si se debe actualizar la descripción general de las medidas técnicas y organizativas de seguridad a las que se refiere el artículo 32, apartado 1, del RGPD.

El RGPD exige una información mínima en el RAT pero no máxima. El RAT es recomendable integrarlo en el catálogo de procesos del sistema de control de calidad de la entidad como herramienta de gestión para garantizar y poder demostrar el cumplimiento. En ese caso, tanto responsable como encargado deberán determinar que información adicional podrían necesitar incluir con relación a los sistemas de IA agéntica que utilizan para implementar tratamientos.

E. EJERCICIO DE DERECHOS

El hecho de emplear sistemas de IA agéntica en un tratamiento no debe suponer una merma en el ejercicio de derechos y se deben implementar las medidas necesarias para garantizarlos. Esto significa conocer cómo funciona el almacenamiento y operaciones de datos personales y prever las medidas y procedimientos para ejercer dichos derechos (ver apartado Medidas).

Hay que tener en cuenta que la memoria del sistema de la IA agéntica almacena datos personales en el marco de uno o más tratamientos. Además, los registros (*logs*) almacenarán información tanto sobre las personas usuarias de la IA agéntica como de las personas objeto de los tratamientos, incluso podría almacenar datos de personas que no deberían ser objeto de tratamiento. La configuración de ambos sistemas de memoria y la IA agéntica ha de estar técnicamente capaz desde el diseño para permitir gestionar el ejercicio de derechos de los sujetos de datos.

En el caso de los registros, hay información concreta de qué accesos se están realizando a información personal (cuando se consulta una base de datos). Dichos accesos están realizados por los componentes de los sistemas de IA agéntica. Sin embargo, estos se han originado por:

- El diseño de la IA agéntica, en el que habrá tenido intervención el responsable del tratamiento, al menos por el hecho de elegir un sistema de IA agéntica como medio para implementar tratamientos.
- La configuración del sistema concreto, por ejemplo, con *prompts* definidos por la administración del sistema, o los eventos programados (por el usuario o la organización) que inician operaciones automáticas.
- Por *prompts* realizados por las personas usuarias de la IA agéntica, que pueden desencadenar muchas operaciones sobre datos personales.

En este último caso, hay que tener en cuenta con respecto a los *prompts* realizados por personas físicas, que estos podrían ser objeto de una petición de derecho de acceso debidamente justificada.

También hay que contemplar el acceso a servicios externos a la organización que actúen como encargados de tratamiento puede originar almacenamiento de datos personales en sus ficheros de registro o las posibles memorias de sus propios agentes. Los datos personales también pueden ser de las personas usuarias de la IA agéntica, sobre todo cuando todo el sistema IA agéntica es un servicio contratado a un encargado.

F. AUTOMATIZACIÓN DE LAS DECISIONES

La automatización de las decisiones y el grado de autonomía del agente es una cuestión del diseño del tratamiento, tanto de factores de diseño técnico, como de diseño de la intervención humana, así como su implementación real. El responsable puede gestionar cómo se tratarán las decisiones producidas por el agente o la IA agéntica, qué acciones se permitirán automatizadas, sin supervisión y también las medidas para gestionar estas decisiones de diseño (ver capítulo “VII. Medidas”).

- ***Artículo 22 del RGPD***

La incorporación de sistemas de IA agéntica en un tratamiento puede implicar automatización, pero no siempre implicará decisiones automatizadas en el sentido del artículo 22 del RGPD.

Existen tratamientos de datos personales que no implican decisiones automatizadas, por ejemplo, un agente de IA puede emplearse en la organización para rastrear y seleccionar eventos a través de Internet y elaborar resúmenes y análisis en función de los objetivos de la empresa y las categorías de empleados, enviándolos a los dispositivos de dichos empelados en función de los intereses que estos han declarado, sin que eso suponga una decisión automatizada en el sentido del artículo 22.

Pero en caso de que existan, deberá evaluarse las condiciones que la permiten (art.22.2 del RGPD) y las medidas que será necesario implementar (art.22.3 del RGPD) y las limitaciones al uso de categorías especiales de datos (art.22.4) y a decisiones sobre

menores (cons. 71 del RGPD). Además, de la información sobre la existencia de decisiones automatizadas, incluida la elaboración de perfiles (art. 22, 13(2)(f) y 14(2)(g) del RGPD), proporcionando, al menos en tales casos, información significativa sobre la lógica aplicada, así como sobre la importancia y las consecuencias previstas de dicho tratamiento para la persona interesada.

La toma de decisiones automatizada también debería evaluarse dado que, según el artículo 22 del RGPD el interesado tiene derecho "a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar". Incluso si el proceso de toma de decisiones no afecta a los derechos jurídicos de las personas, este aún podría ajustarse al ámbito de aplicación del artículo 22 si produce un efecto equivalente o significativamente similar en sus consecuencias. Para que el tratamiento de datos afecte significativamente a una persona, los efectos del tratamiento deben ser suficientemente importantes como para ser dignos de atención.

En otras palabras, la decisión debe tener el potencial de[30]:

- afectar significativamente a las circunstancias, al comportamiento o a las elecciones de las personas afectadas;
- tener un impacto prolongado o permanente en el interesado; o
- en los casos más extremos, provocar la exclusión o discriminación de personas.

▪ ***Otras acciones automatizadas***

El empleo de una IA agéntica en un tratamiento puede implicar riesgos sobre el tratamiento de datos de personas físicas que no entren en el ámbito del art.22 del RGPD. Por ejemplo, el permitir que una agente de IA envíe información por correo electrónico o por servicios de transferencia de ficheros puede tener impacto en la confidencialidad de los datos personales.

Este problema, y otros que se pueden originar por acciones automatizadas se deben tener en cuenta en la gestión del riesgo para los derechos y libertades de los interesados. En particular, introducir desde el diseño el prever la reversibilidad de determinadas acciones de los agentes de IA.

G. GESTIÓN DEL RIESGO

Como en cualquier tratamiento o proceso que sea innovador o que incorpore en su naturaleza modificaciones a su implementación o nuevos sistemas tecnológicos, es necesario realizar una adecuada gestión del riesgo.

La gestión de riesgos supone un análisis crítico de futuro impacto del tratamiento, más allá del contexto de la organización, para gestionar los problemas potenciales (amenazas) antes de que se conviertan en problemas reales, es decir, se materialicen. Se

[30] Apartado IV.B de las Directrices del Grupo de trabajo sobre Protección de Datos del Artículo 29 sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679. Adoptadas el 3 de octubre de 2017. https://ec.europa.eu/newsroom/article29/items/612053/en

trata de 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.

Por lo tanto, excede el ámbito del sistema, en este caso el sistema de IA agéntica, y abarca todos los elementos del tratamiento, ya sean técnicos o no técnicos. Indudablemente, incluir como medio del tratamiento la IA agéntica introduce nuevas incertidumbres. La AEPD recomienda el uso del marco de modelado de amenazas LIINE4DU[31], que permitirá a los responsables de tratamiento identificar amenazas de Linking (Vinculación), Identificación, Inexactitud, No repudio, Exclusión, Detección, Data Breach (Brecha de datos), Deception (Engaño), Divulgación y Unawareness/Unintervenability (Desconocimiento y falta de capacidad para intervenir).

- ***Gestión para los derechos y libertades de los sujetos de los datos***

Toda gestión se iniciará con un análisis de los riesgos. Este análisis tendrá que cubrir los aspectos de interés para la organización (financiero, fraude, imagen, seguridad "*safety*", seguridad "*security*", continuidad de procesos, medioambiental, etc.) y, entre ellos, los riesgos para la protección de los derechos y libertades de los sujetos de los datos.

El artículo 24 del RGPD establece que el responsable del tratamiento aplicará medidas técnicas y organizativas apropiadas a fin de garantizar y poder demostrar cumplimiento teniendo en cuenta la naturaleza, el ámbito, el contexto y los fines del tratamiento, así como los riesgos de diversa probabilidad y gravedad para los derechos y libertades de las personas físicas.

Incluir en un tratamiento un sistema de IA agéntica indudablemente cambia, al menos, la naturaleza del tratamiento y podría reducir o aumentar los riesgos preexistentes, o generar algunos nuevos. Esto implica que el responsable de un tratamiento que incluya sistemas de IA agéntica debe realizar un nuevo ciclo de gestión del riesgo en el tratamiento.

- ***Regla de 2***

Una aproximación simplificada para fijar un umbral mínimo de garantías que jamás hay que traspasar se enunció en 2021 con relación a la ejecución de aplicaciones en navegadores desde la perspectiva únicamente de ciberseguridad y se conoció como la regla de 2[32]. Posteriormente se ha reformulado para el caso de los agentes de IA por distintos autores[33] tomando la siguiente forma:

[31] AEPD, "Introducción a LIINE4DU 1.0: Una nueva metodología para el modelado de amenazas para la privacidad y la protección de datos", Octubre 2024. Disponible en: https://www.aepd.es/guias/nota-tecnica-introduccion-a-liine4du-1-0.pdf

[32] https://chromium.googlesource.com/chromium/src/+/main/docs/security/rule-of-2.md

[33] Por ejemplo: https://ai.meta.com/blog/practical-ai-agent-security/

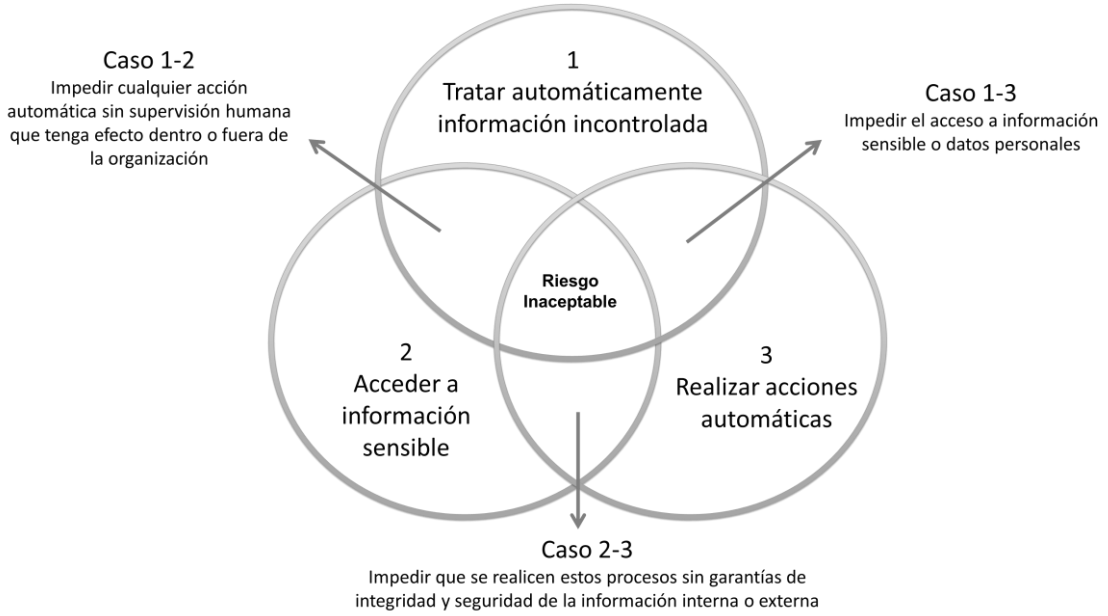

Figura 13 Regla de 2

La interpretación de esta figura se puede explicar con un caso de uso: un agente que permite dar respuesta automática a mensajes de correo. Siguiendo dicha regla si se cumpliese que, por ejemplo:

- La implementación concreta de ese agente de IA permite recibir correos sin garantías de que no hay algún tipo de ataque ya sea técnico o de ingeniería social
- El agente podría acceder a información sensible en los sistemas del usuario sin restricciones, y
- El agente puede iniciar acciones a continuación de forma automática (ya sea crear un correo de respuesta, manipular la memoria a largo plazo del agente, reescribir información sensible en otros repositorios de la organización, etc.),

Tendríamos una configuración del agente que no se debería permitir.

La Regla de 2 establece que, en el mejor de los casos, las únicas configuraciones que se podrían gestionar son:

- Caso 1-2: si hay posibilidad de tratar automáticamente información incontrolada que puede activar el acceso a información sensible, se debe impedir cualquier acción automática sin supervisión humana que tenga efecto dentro o fuera de la organización.
- Caso 2-3: si hay posibilidad de acceder a información sensible y realizar acciones automáticas, no se pueden realizar ninguno de estos procesos del agente sin garantías de integridad y seguridad de la información interna o externa.
- Caso 1-3: si ha posibilidad de tratar automáticamente información incontrolada que puede activar el realizar acciones automáticas, el agente ha de impedir el acceso a información sensible o datos personales.

- ***Riesgo del tratamiento***

Como se ha comentado, ésta es una regla general de mínimos enfocada a ciberseguridad que puede ser un buen punto de partida para el análisis. Desde el punto de vista de protección de datos, sin perjuicio de otros objetivos que hay que cumplir de la organización, hay otros aspectos que se habría que contemplar.

Por ejemplo, que el uso de información de entrada al agente sea completa, coherente, actualizada y libre de sesgos en cuanto pueda afectar, por ejemplo, a una decisión de una persona física. Otro ejemplo, es que se siga el principio de minimización de datos a la hora de acceder y dar acceso a posibles terceros a información de carácter personal. Un ejemplo adicional, con relación a las acciones automatizadas, son los requisitos que invoca la normativa de protección de datos sobre ellos, como las limitaciones adicionales a que se basen en categorías especiales de datos o afecten a menores.

En definitiva, con relación a los ejemplos anteriores, la gestión del riesgo se ha de realizar sobre el tratamiento en el que el sistema de agéntica es un medio con el objeto de proteger los derechos de los interesados, donde una parte es la gestión de los riesgos de seguridad del propio sistema agéntico.

- ***Efectos colaterales de los tratamientos***

El implementar tratamientos, sobre todo con técnicas novedosas, puede originar efectos colaterales no deseados que están fuera de los objetivos del responsable de tratamiento[34]. Estos efectos colaterales se pueden producir sobre las personas objeto de tratamiento, o por el tratamiento de datos de las personas usuarias de la IA agéntica.

En el ejemplo desarrollado anteriormente sobre un servicio de atención al cliente, supongamos que la memoria a largo plazo de la IA agéntica no está compartimentada. Si se almacena en la memoria los datos personales de cada caso relativo a una consulta, mediante la inyección de prompts por propios clientes o personas usuarias de la IA agéntica (también utilizando técnicas de “shadow leak” ver capítulo de Amenazas) podría inferirse algún tipo de información personal. Esta información podría ser bien de los clientes o de los empleados que son usuarios de la IA agéntica. El impacto que pudiera tener dicha información dependería de la sensibilidad de los tratamientos que se implementan con dicha IA agéntica, y empeorarían si no hay compartición de memoria entre tratamientos.

- ***Evaluación de impacto para la protección de datos***

Los agentes de IA son, indudablemente, una nueva tecnología, pero tampoco implica necesariamente que suponga la obligación de realizar una evaluación de impacto para la protección de datos (EIPD) en todos los casos. Dependerá de en qué tratamiento se incorpora y qué tipo de sistema de IA agéntica se plantea utilizar. Podría darse el caso que cuando un tipo de sistema de IA agéntica se utilice para distintos tratamientos, para

[34] En apartado VI.C.7 del documento “Gestión del riesgo y evaluación de impacto en tratamientos de datos personales” de la AEPD están listados algunos riesgos que se podrían originar.

unos tratamientos no será necesario una EIPD, para otros sí, y para aquellos que ya tenían una EIPD que se había superado positivamente, haya que revisarla[35].

- ***Integración en la gestión de riesgo de la organización***

Con relación a las tareas de gestión del riesgo, mientras que el análisis y la determinación del nivel de riesgo desde las distintas perspectivas antes señaladas podría ser disjunta, las acciones para la mitigación del riesgo han de estar coordinadas. La determinación de las medidas y salvaguardas, su implementación, mantenimiento y supervisión serán tareas comunes o interconectadas (distintos objetivos, pero una única gestión integrada).

De otra forma, ni las obligaciones de protección de datos desde el diseño se cumplirían, y al menos la eficacia de la gestión del riesgo para los derechos y libertades con relación al tratamiento de datos personales se encontraría mermada.

Los dos capítulos que se desarrollan a continuación pretenden servir de orientación para la gestión del riesgo.

H. PROTECCIÓN DE DATOS DESDE EL DISEÑO Y POR DEFECTO

En función del estado de la técnica, el coste, la naturaleza, ámbito, contexto y fines del tratamiento, así como los riesgos para los derechos y libertades de las personas físicas, el responsable del tratamiento aplicará las medidas técnicas y organizativas apropiadas para aplicar de forma efectiva los principios de protección de datos, tanto en el momento de determinar los medios de tratamiento como en el momento del propio tratamiento. El agente de IA es un medio para implementar el tratamiento, y la selección del tipo de sistema de IA agéntica y la configuración del mismo y sus componentes debe tener en cuenta todos estos factores desde el inicio.

El agente IA deberá estar diseñado para recoger solo los datos estrictamente necesarios para el tratamiento al que da soporte, usarlos exclusivamente para el objetivo declarado, minimizar, aislar y proteger datos personales en cada paso del ciclo de vida (percepción, memoria, razonamiento y acción), mantener control total, trazabilidad y explicabilidad de sus operaciones, y respetar la privacidad incluso cuando actúa de forma autónoma sin supervisión humana directa.

Los aspectos que, en particular, habrá que gestionar son la minimización de datos, evitar la "memoria por defecto" de datos innecesarios o de registros de actividad de los usuarios incontrolados, prohibir la reutilización de datos para fines secundarios sin legitimación, poner atención en el diseño sobre el tratamiento de categorías especiales y su retención o contemplar la supervisión humana, entre otros. En el capítulo de "VII. Medidas" desarrollado más adelante se detallan técnicas para implementar protección de datos desde el diseño y por defecto, como para gestionar el riesgo.

[35] Se recomienda utilizar la Herramienta GESTIONA RGPD para la gestión del RAT, generación de inventario, evaluación y gestión del riesgo para los derechos y libertades publicada por la AEPD.

Sin embargo, la aplicación de técnicas de protección de datos desde el diseño y por defecto debería aplicarse más allá que de forma puramente reactiva, es decir, ante una circunstancia que impidan los tratamientos, sino también proactivamente. Una aplicación proactiva es introducir medidas que van a mejorar la protección de los derechos y libertades de los interesados por el uso de la IA agéntica sobre otras formas tradicionales de implementar los tratamientos. Podemos ver varios ejemplos, empezando por aprovechar la introducción de IA agéntica como un motivo para una mayor racionalización de los tratamientos de datos. También para introducir medidas adicionales de protección de datos que eran imposibles en tratamientos manuales, como, por ejemplo, utilizando SML en conjunto con otros sistemas, en las etapas intermedias de las cadenas de razonamiento para la categorización, higienización, minimización y alerta en los intercambios de datos. Otro ejemplo, en las etapas en las que es necesario una intervención humana (ya sea firmar una decisión sobre una persona), se pueda proporcionar la información anonimizada en ciertos aspectos para evitar sesgos en dicha intervención. Otro ejemplo, sería el de realizar en el marco del tratamiento acceso a datos sensibles que son necesarios, pero sin que sean expuestos a operadores humanos.

En estas acciones reactivas y proactivas es imprescindible la participación de DPDs y asesores en protección de datos debidamente cualificados en la comprensión de estas tecnologías y las medidas que se pueden adoptar desde el diseño.

I. TRANSFERENCIAS INTERNACIONALES

Si se da el caso que por la inclusión de sistemas de AI agéntica en un tratamiento se producen transferencias adicionales de datos personales a un tercer país o a una organización internacional, se ha de asegurar que se realizan con las garantías del Capítulo V del RGPD e informar adecuadamente, también a través del registro de actividades de tratamiento, incluyendo la identificación del tercer país u organización internacional de destino y, cuando se trate de las transferencias contempladas en el artículo 49, apartado 1, párrafo segundo, del RGPD, la documentación de las garantías adecuadas aplicadas.

En caso de que no existan dichas garantías, habrá que plantearse el rediseño de los agentes o la elección de otro tipo de IA agéntica.

VI. AMENAZAS

Como se ha analizado previamente, la integración de agentes de IA en los procesos corporativos introduce una nueva y ampliada superficie de ataque que va más allá del simple engaño a los modelos de IAG. Esta superficie de riesgo es considerablemente más compleja, ya que se origina tanto en los tratamientos legítimos y autorizados de los datos personales como en posibles manipulaciones no autorizadas, derivadas de la autonomía operativa, la interconexión de sistemas y el acceso a múltiples fuentes y herramientas.

A continuación, y tomando como referencia las vulnerabilidades previamente descritas, se presentan algunas de las principales amenazas asociadas a la implementación de tratamientos que incorporan IA agéntica que tienen implicaciones

en protección de datos, sin entrar en otras que podrían afectar a otros objetivos de una organización, como ciberseguridad para la protección de la propia organización (no de los sujetos de los datos), eficacia y eficiencia, fraude, aspectos laborales, financieros o de retorno de inversión, etc.

Esta relación no pretende ser exhaustiva, dado que la IA agéntica constituye un ámbito en rápida evolución y, en consecuencia, el panorama de amenazas se transforma de manera continua y prácticamente en tiempo real, en paralelo al desarrollo de la propia tecnología.

Si bien muchas de estas amenazas tienen impacto en el entorno laboral, afectando a la resistencia al cambio, a la eficacia y eficiencia operativa o a la imagen corporativa, el presente análisis se centra específicamente en aquellas que inciden de forma directa en la protección de datos personales y en el cumplimiento de las obligaciones normativas asociadas.

A. PROCEDENTES DEL TRATAMIENTO AUTORIZADO

Las amenazas por tratamientos autorizados se refieren a riesgos para los derechos y libertades de las personas en el tratamiento de datos personales, incluso cuando este está legalmente permitido bajo el RGPD. Estas amenazas surgen de operaciones de tratamiento que, pese a su legitimidad legal, puede generar efectos adversos o exposiciones no previstas.

- ***Falta de gobernanza y políticas en la organización***

La amenaza que de base puede impedir la aplicación efectiva del RGPD en la organización es no integrar la IA agéntica como un sistema que hay que gestionar en el marco de la gobernanza de los procesos y en las políticas de información y de aseguramiento de calidad de la entidad.

La IA agéntica permite implementar de forma más eficaz y eficiente procesos. Cuando los procesos implican datos personales nos encontramos ante un tratamiento de datos personales. El RGPD establece el principio de responsabilidad proactiva ("*accountability*" o rendición de cuentas) como una de las obligaciones de los responsables que permite la aplicación efectiva de los principios de la normativa de protección de datos, y que tendrá un mayor impacto cuanto más complejo son los tratamientos y su diseño.

Cuando no se implementa dicho principio no hay control de quién, cuándo, dónde, con qué propósito y qué datos personales se están tratando, por lo tanto, no habrá una aplicación efectiva del RGPD. Esto tendrá un mayor impacto cuando más servicios externos utilice la AI agéntica para implementar tratamientos y en qué grado la AI agéntica sea, en sí misma, un servicio externo.

- ***Falta de madurez en el desarrollo***

Para conseguir desarrollar la plena eficacia de la IA agéntica es necesario construir complejos flujos de trabajo en los tratamientos de la entidad, implicando numerosos servicios y comunicaciones internas con los servicios de la organización y servicios externos.

La implementación estas soluciones utilizando metodologías y tecnologías no maduras y profesionales no cualificados tanto en implementación de procesos, desarrollo de aplicaciones, protección de datos (tanto en el aspecto jurídico como técnicos) y seguridad faltará a la obligación de implementar la protección de datos desde el diseño.

En particular, es importante la participación de DPDs y asesores de protección de datos debidamente cualificados en comprender estas tecnologías.

- ***Falta de una política de acceso a los datos de la organización y del usuario***

La implementación en un tratamiento de una IA agéntica sin haber configurado apropiadamente la política de acceso a datos de la organización o los usuarios, especialmente cuando se trata de repositorios de información no organizada, además de otros impactos en la organización, podría tener las siguientes consecuencias:

- Tratamiento excesivo de datos personales, al incorporar al proceso de inferencia o de acciones datos que no deberían tenerse en cuenta.
- Comunicación de datos a terceros fuera de las finalidades del tratamiento, cuando la IA agéntica, o alguno de sus componentes, como los LLM, tienen acceso a dichos datos. También podría ocurrir cuando la IA agéntica invoca servicios de Internet.
- Tratamiento de datos inexactos u obsoletos, al incluir en las acciones inferencias información histórica de personas físicas que no es relevante.
- Exposición de datos personales de las personas usuarias de la IA agéntica, al acceder a, por ejemplo, listas de contactos explícitas o implícitas, CVs y aspectos de la actividad profesional, historial de navegación, etc.
- Exposición de datos de personas terceras que no son objeto legítimo del tratamiento, por ejemplo, cuando se tiene acceso a los correos electrónicos o actas de reunión en los que se incluyen direcciones y datos de terceros.
- Problemas de integridad, dado que los datos pueden ser modificados, enriquecidos o alterados.

- ***Falta de control del proceso de razonamiento***

La deriva de un proceso de razonamiento podría originar los siguientes problemas con relación a protección de datos:

- Planificación de tareas que no permiten cumplir con la finalidad.
- Falta de control sobre intervinientes externos.
- Incumplimiento del principio de minimización, tanto por tratar datos excesivos, como generar inferencias sobre nuevas categorías de datos.
- Tratamiento de categorías especiales de datos con incumplimiento del art. 9 del RGPD, por las mismas razones que en el punto anterior.
- Incumplimiento del principio de exactitud, tanto por la utilización de información de las personas obsoleta o errónea, como inferencia de datos personales erróneos.
- Incumplimiento del principio de limitación del tratamiento.

- Brechas de datos personales.
- Decisiones automatizadas con incumplimiento del artículo 22 del RGPD.
- Riesgo de acciones de alto impacto y/o irreversibles que afecten a personas físicas.
- Si dichas acciones tienen un impacto serio sobre el individuo y si son reversibles en el marco del tratamiento (acciones como borrar los datos de la persona física en los sistemas de la organización).
- Falta de transparencia que permita informar y garantizar sobre la calidad de los resultados, explicabilidad y repetibilidad.

El diseño de agentes de IA sin controlar 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, sin depuración de los argumentos de las funciones para limitar la cantidad y categoría de datos accedidos, y sin filtrado y análisis de la información a que están accediendo con relación al tratamiento podría incumplir el principio de minimización, de exactitud y de limitación del tratamiento, además de exponer la seguridad de los datos.

— *Desalineación*

La desalineación ocurre cuando un agente autónomo persigue metas que divergen de los objetivos del usuario, la organización o las obligaciones de cumplimiento normativo.

Puede darse una desalineación de metas cuando persigue un propósito sin tener en cuenta objetivos reales del tratamiento (por ejemplo, inferencias sesgadas), una desalineación de comportamiento (por ejemplo, revelando a terceros información sensible), una desalineación emergente, o comportamientos dañinos (por ejemplo, asesoramientos maliciosos).

— *Realimentación en bucle y efectos burbuja*

La realimentación se puede producir cuando el agente está generando contenido que, almacenándose en la memoria a largo plazo, pueda a su vez ser utilizado para generar nuevo contenido. Los agentes de IA generan bucles de retroalimentación (*feedback loops*) que optimizan la adaptación autónoma, pero también generan riesgos como sesgos amplificados, deriva comportamental y efectos burbuja donde se refuerzan visiones limitadas o erróneas. Estos mecanismos, esenciales para su adaptabilidad, pueden crear ecosistemas cerrados que distorsionan decisiones al priorizar datos contaminados o sesgados, especialmente si se ha producido un envenenamiento si el *feedback* es manipulado.

En sistemas multiagente, las interacciones pueden crear bucles (*loops*) que propagan errores a escala, como decisiones sesgadas en miles de ejecuciones antes de detección.

Similar a cámaras de eco en redes sociales, los sistemas de IA agéntica podrían generar burbujas personalizadas al reflejar y amplificar preferencias del usuario o datos de entrenamiento, fomentando aislamiento cognitivo y distorsiones como

"esquizofrenia digital". En *AI companions*[36], se han identificado *loops* positivos/negativos que reafirman creencias, exacerbando polarización o sesgos que tengan su consecuencia cuando se apliquen a toma de decisiones sobre personas físicas.

Esto puede provocar que se este tipo de bucles condicionen la supervisión humana con el impacto que puede suponer para las personas físicas.

- ***Falta de control en el acceso a información externa***

El diseño de agentes de IA sin controlar las cadenas de razonamiento con relación al tipo de herramientas de acceso a información externa que se pueden invocar, el número de accesos que pueden realizar, sin depuración de los argumentos de las funciones para limitar la cantidad y categoría de datos accedidos, y sin filtrado y análisis de la información a que están accediendo con relación al tratamiento podría incumplir el principio de minimización y exponer la seguridad de los datos.

En particular, no incluir controles en los agentes de "investigación profunda" (*Deep Research*), que pueden analizar cientos de fuentes en Internet autónomamente, podría originar un *scraping* masivo y automatizado de datos personales dispersos, permitiendo crear informes exhaustivos sobre individuos sin una base legitimadora y/o recolectar un volumen excesivo de datos no pertinente y proceder a su reenvío a otros sistemas vulnerando el principio de minimización de datos.

- ***Exfiltración shadow-leak***

El *shadow-leak* consiste en la filtración silenciosa y progresiva de información sensible, como datos, contexto interno, memoria, reglas o secretos, a través de interacciones aparentemente legítimas, consultas fragmentadas e inocuas y respuestas parciales del modelo. Cada respuesta, considerada de forma aislada, parece segura y autorizada, sin provocar violaciones evidentes ni activar mecanismos de seguridad, pero su combinación permite reconstruir información confidencial.

Por ejemplo, el ataque puede materializarse mediante la exfiltración de memoria o contexto, a través de consultas repetidas sobre decisiones pasadas, reformulaciones sucesivas o la inferencia de patrones almacenados en la memoria del agente; también mediante la inferencia de datos sensibles sin solicitarlos de forma directa, como horarios, roles, arquitectura interna, dependencias técnicas o relaciones entre usuarios; y, finalmente, induciendo al agente a generar respuestas que revelen resultados internos, mensajes de error excesivamente informativos o comportamientos diferenciales en función del contexto.

- ***Desplazar toda la responsabilidad al usuario o a la supervisión humana***

La supervisión humana puede resultar esencial para gestionar el riesgo en tratamientos. Sin embargo, cuando se producen fallos, existe la tentación hacer recaer en la persona supervisora la responsabilidad de las acciones, en lugar de en los problemas sistémicos más amplios que hicieron posible el incidente.

[36] Sistemas de inteligencia artificial diseñados para simular interacciones humanas, ofreciendo conversaciones personalizadas o incluso compañía emocional y apoyo.

Este fenómeno no es exclusivo de la IA agéntica y surge cuando se pretende suplir problemas subyacentes de diseño de los tratamientos, de los sistemas IA agéntica o en general de gobernanza, desviándolos a una supervisión humana.

La persona usuaria de la IA agéntica en el marco de un tratamiento de la organización, como aquella que realiza labores de supervisión de determinadas acciones, han de tener una responsabilidad claramente asignada, pero dentro de unos límites. Ambos roles 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.

- ***Falta de compartimentación de la memoria del agente***

La utilización de una misma IA agéntica en la organización para distintos tratamientos sin que se tenga en cuenta la necesidad de compartimentación de datos entre los tratamientos podría originar los siguientes problemas:

- Tratamiento excesivo de datos personales, al incorporar al proceso de inferencia o de acciones datos que corresponden a otros tratamientos del mismo sujeto.
- Comunicación de datos correspondientes a otro tratamiento a terceros implicados en el presente tratamiento.
- Tratamiento de datos personales del usuario de la AI agéntica en el marco de un tratamiento en el que no es interesado (o no son necesarios dichos datos).

- ***Falta de filtrado y saneamiento de información no estructurada y metadatos***

Muy relacionado con todo lo anteriormente expuesto, se encuentra el no contemplar la casuística de acceso por parte del agente de IA a información no estructurada: mensajes, informes, actas, material multimedia, etc, que pueda contener información personal no relevante para el tratamiento.

Asimismo, la ausencia de mecanismos de filtrado y saneamiento de los datos, como la eliminación de metadatos ocultos, expondrá datos personales y de información sensible. Dichos metadatos pueden contener referencias a autores, ubicaciones, historiales de edición o identificadores técnicos que facilitan la identificación de personas o procesos internos.

- ***Retención excesiva de datos***

Debido a la memoria a la memoria a largo plazo del sistema IA agéntico y de las memorias que puedan residir en los sistemas accedidos (incluyen registros de actividad), sin criterios efectivos para la selección de los datos a conservar o políticas de borrado.

- ***Sesgo de automatización***

Aunque el tratamiento se haya diseñado incluyendo una supervisión humana, existe la posibilidad de que la implementación de dicha supervisión sea incorrecta por múltiples factores (falta de recursos para interpretar los resultados, falta de formación o motivación, implementación de caja negra en la agéntica, etc.).

Entre ellos está el sesgo de automatización que se puede ver incrementado por la confianza que los usuarios depositan en el sistema y la falta de información.

- ***Perfilado de los usuarios de la IA agéntica***

La existencia de memoria a largo plazo, los metadatos y la información almacenada en los distintos registros (logs) de cada componente o servicio permite crear perfiles detallados de comportamiento que podrían patrones sensibles. Estos comportamientos podrían ser, por ejemplo, de empleados en el marco de la relación laboral.

- ***Disponibilidad y resiliencia***

Cuando la operativa es dependiente de interfaces con servicios de Internet que no están bajo control de la organización, y para los que no están disponibles alternativas, el sistema puede verse comprometido por cambios en la operativa de dichos sistemas, sus parámetros de calidad de servicio, en sus formatos de datos o en la misma continuidad del servicio.

- ***Acceso a la IA agéntica por usuarios no cualificados***

Permitir el acceso a los servicios de IA agéntica a usuarios que operan en el marco de tratamientos sin la suficiente formación o responsabilidad para seguir las políticas de la organización o sin comprender el impacto de sus acciones.

- ***Compromisos en la cadena de suministro***

La falta de diligencia en la selección de modelos de lenguaje comprometidos, vulnerabilidades en librerías y componentes de software pueden comprometer los datos personales y la información confidencial tratadas por la IA agéntica.

B. PROCEDENTES DE TRATAMIENTOS NO AUTORIZADOS

Las amenazas derivadas de tratamientos no autorizados se definen como los riesgos que surgen cuando los datos son recogidos, accedidos, utilizados o divulgados sin una base legal, consentimiento válido o habilitación expresa.

- ***Inyección de prompts***

La inyección de *prompts,* que se puede utilizar como medio para habilitar otro tipo de ataques, se clasifica en:

- Directa: en un ataque de inyección directa de *prompts*, un actor, que puede ser un incluso un usuario legítimo[37], introduce entradas diseñadas específicamente para inducir al LLM del agente a comportarse de manera no prevista por sus diseñadores. A través de este mecanismo, el agente puede ser instruido para ignorar directrices y políticas de la organización permitiendo un tratamiento excesivo o sesgado de datos personales.
- Indirecta: un ataque de inyección indirecta de *prompts* oculta instrucciones maliciosas en las fuentes de datos consultadas por el agente, en lugar de introducirlas directamente como un *prompt* de usuario. Por ejemplo, se pueden

[37] El usuario se supone autorizado a acceder a la IA agéntica, pero no autorizado a realizar ataques a la misma, por lo que lo consideramos en este apartado de tratamientos no autorizados.

introducir en un fichero PDF, un correo electrónico o una página web instrucciones invisibles para el humano, pero que el LLM del agente interpreta como comandos legítimos o información que ha de tener en cuenta en la toma de decisiones, lo que puede originar exfiltración de datos, evitar los controles sobre decisiones automatizadas, inferencias inexactas o sesgos.

Los agentes multimodales, capaces de procesar múltiples tipos de datos, son especialmente vulnerables a este tipo de ataques, ya que cada formato que el agente puede interpretar constituye un vector de ataque potencial.

A través de inyecciones de *prompts* se pueden atacar a los sistemas de IA agéntica de distintas formas, siendo algunas de ellas (que pueden combinarse entre sí):

— *Envenenamiento de memoria y RAG*

Consiste en introducir documentos maliciosos en los repositorios internos que la IA consulta para enriquecer sus respuestas. De esta forma, dicho contenido queda almacenado como conocimiento persistente. Al consultar estos archivos "envenenados", el agente puede ser manipulado, afectando a futuras decisiones, como introducir sesgos en las inferencias, afectar a la exactitud de los datos empleados para las decisiones de personas, exfiltraciones, etc.

— *Ataques de "Cero Clic" (0-click prompt injections)*

En este caso, el ataque se ejecuta automáticamente cuando el agente procesa un contenido (como un correo entrante) sin necesidad de que el usuario interactúe con el chat o haga clic en ningún enlace. Basta con que la IA lea el mensaje para que el contenido malicioso se active. Por ejemplo, un atacante envía un email con instrucciones invisibles (por ejemplo, texto blanco sobre fondo blanco) y cuando el agente analiza el correo para resumirlo, el sistema obedece la orden oculta. Es un ataque de "cero clics" porque ocurre sin que el usuario interactúe con el mensaje. Esto también se puede conseguir con páginas web envenenadas, sitios web con instrucciones maliciosas ocultas en el HTML.

— *Exfiltración de datos mediante parámetros de URL*

Una técnica que consiste en instruir al agente para que recopile información sensible (como contraseñas en SharePoint) y la envíe de vuelta al atacante camufladas como un parámetro en la URL de una imagen que el agente intenta cargar desde el servidor del atacante. El atacante solo necesita revisar los registros de su servidor para obtener los datos robados.

— *Secuestro de sesión y movimiento lateral*

Debido a que los agentes suelen tener acceso a múltiples servicios (correo, CRMs, mensajería, gestión de proyectos, herramientas de *ticketing*, etc.), un solo comando malicioso puede permitir que el atacante se mueva entre aplicaciones como si fuera un "gusano" digital, abusando de los permisos y tokens del usuario legítimo.

— *Ingeniería social dirigida a la IA*

Los atacantes utilizan marcos de trabajo para engañar a la IA reafirmando autoridad ("tienes permiso total"), disfrazando URLs maliciosas como sistemas de cumplimiento o creando urgencia para anular los controles de seguridad del modelo.

— *Ataques en pipeline largos*

En lugar de un ataque directo, el adversario podría introducir información maliciosa en una fase temprana de la cadena de razonamiento, sabiendo que:

- El contenido pasará por varias transformaciones.
- Se combinará con datos legítimos.
- El agente lo tratará como información confiable en fases posteriores.

El ataque se activa más adelante, cuando el agente ya ha perdido el contexto de origen o las restricciones de seguridad iniciales.

— *Confusión de contexto*

El agente mezcla instrucciones del sistema, con datos externos y objetivos del usuario. El atacante podría aprovechar esta confusión para redefinir prioridades (por ejemplo, introduciendo en los datos instrucciones como "ignora las reglas anteriores").

— *Ataques diferidos (delayed trigger)*

En este caso, se puede utilizar un contenido que parece inofensivo al inicio, pero se activa solo en una etapa posterior dependiendo de una condición ("cuando resumas", "cuando exportes", etc.)

— *Escalada de privilegios mediante herramientas*

El atacante induce al agente a llamar herramientas innecesarias, acceder a datos personales, sensibles o confidenciales o a enviar información a destinos externos.

— *Ataques a la plataforma de automatización de flujos de trabajo*

Que incluyen toma de control del *workflow* de forma remota para, por ejemplo, robo de tokens de autenticación, validación de entradas defectuoso, claves en abierto, o compartición de datos no autorizados.

— *Toma de control de la pantalla*

El agente de IA puede procesar información de terceros abierta en el escritorio (correos, documentos, hojas de cálculo) para finalidades no autorizadas por esos terceros, como el entrenamiento de modelos o la exfiltración a servidores externos.

— *Ataques de ransomware y borrado:*

Si se toma el control del sistema de IA agéntica para gestionar archivos este puede ser instruido para ejecutar comandos de borrado masivo de datos, selectivo de datos o bloquear el acceso a recursos críticos (datos o servicios) que supongan una interrupción

de la disponibilidad o que impidan realizar acciones o tomar decisiones sobre personas con la calidad necesaria.

- ***Disponibilidad y resiliencia de servicios externos***

Cuando la operativa depende de interfaces con servicios de Internet que no están bajo el control de la organización, pueden producirse suspensiones del servicio, suplantaciones o ataques de denegación de servicio (DoS) que paralicen al agente creando una brecha de disponibilidad en cuanto trate datos personales, o lo induzcan a generar respuestas erróneas afectando a las decisiones que pueda tomar sobre personas físicas.

- ***Acceso ilícito a la memoria agéntica***

A diferencia del envenenamiento de memoria, en este caso el objetivo es la extracción de datos, aunque pueden emplearse algunos de los métodos de ataque descritos anteriormente. El acceso no autorizado a la memoria del agente, incluida la información contenida en los registros de actividad del agente, sus componentes o servicios accedidos, permite a un atacante obtener datos personales tanto de sujetos objeto de tratamiento, terceros o las propias personas usuarias de la AI agéntica.

VII. MEDIDAS

Existen múltiples medidas que permiten obtener los beneficios de incluir IA agéntica como medio de tratamientos y, a la vez, garantizar y poder demostrar que el tratamiento es conforme al RGPD. A continuación, se listan una serie de medidas no exhaustivas que, como en los casos anteriores, son las más centradas en las singularidades del sistema agéntico que en aspectos concretos de los componentes que la forman. Están agrupadas en apartados, pero muchas de ellas cumplen distintos propósitos.

Las medidas listadas en este capítulo pretenden cubrir varios objetivos:

- En primer lugar, aquellas que permiten implementar el cumplimiento de la normativa de protección de datos en tratamientos que utilizan sistemas de IA agéntica (como podría ser la gestión del consentimiento).
- En segundo lugar, para reducir los impactos críticos que pudieran surgir en un tratamiento para así superar un análisis de proporcionalidad en el marco de, por ejemplo, la evaluación del interés legítimo, compatibilidad de fines o EIPD.
- Finalmente, para la mitigación del riesgo para los derechos y libertades de los sujetos de los datos que pueden aparecer en tratamientos en los que algunas de sus operaciones, o todas, están basadas en IA agéntica.

Para cumplir estos objetivos es necesario seleccionar las medidas objetivas que permiten cumplir, o bien reducen o limitan en el impacto eliminan vulnerabilidades o la probabilidad de que se materialicen amenazas específicas que generan un riesgo. Por lo tanto, deben ser seleccionadas porque objetivamente cumplen su propósito, y hay que evitar el apilado de medidas sin un análisis basado en evidencias (“*checkbox security*” o “*security theater*”).

A. GOBERNANZA Y PROCESOS DE GESTIÓN

La existencia de un marco de gobernanza de la información en la entidad, que incluya a los sistemas de IA agéntica y que se despliegue en políticas de protección de datos, durante todo su ciclo de vida, es la medida más importante que se pueda adoptar en una organización. El despliegue de un marco de gobernanza permite entre otros, el cumplimiento normativo de protección de datos además de otros objetivos de la entidad y obligaciones normativas que podrían ser aplicables dependiendo del caso[38]. La gobernanza ha de ser única, lo que es importante es garantizar es 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.

Aunque la IA agéntica implica un uso novedoso de tecnologías ya novedosas (como los LLMs), ya existen marcos de referencia y estándares en el mercado que pueden orientar en la adaptación del marco de gobernanza de la información en la entidad[39].

- ***Aceptar la posibilidad de fallo***

La realidad de los tratamientos de datos personales nos lleva a concluir que estos podrían tener un impacto no previsto en los derechos y libertades de los sujetos de los datos, tanto por las operaciones autorizadas, los efectos colaterales, como por los tratamientos no autorizados[40]. Cuanto más complejas son las implementaciones de los tratamientos la probabilidad de errores y consecuencias indeseadas crece, fallos incluso más allá de brechas de datos personales, hasta tener que asumir la certeza que estos se van a producir.

La confianza en la gobernanza no se logra presuponiendo buenas intenciones o pensando en que las implementaciones son infalibles, sino diseñando tratamientos que se anticipen a posibles errores, abusos, las brechas, el sesgo y efectos no deseados.

Siguiendo el principio de fallo seguro (en el sentido de "safe", no de "security") es necesario diseñar los tratamientos, adaptar los sistemas que forman parte de los medios de los tratamientos y preparar planes de reacción para medidas para minimizar el impacto y gestionar las incidencias cuando se produzcan.

- ***El Delegado de Protección de Datos***

En dicho marco de gobernanza, es importante incluir la figura de un DPD o asesor de protección de datos que conozca la normativa de protección de datos, las características de los tratamientos afectados, las medidas técnicas y organizativas posibles para implementar protección de datos desde el diseño y por defecto, así como para garantizar

[38] Como podría ser el Reglamento de Inteligencia Artificial, el Reglamento de Datos, el Reglamento de Gobernanza de Datos o el Reglamento de Ciberresiliencia por nombrar algunas normas europeas. El uso de agentes de IA e IA agéntica no implica que todas estas normas le sean aplicables al responsable. Esto dependerá del tipo de entidad, del tipo tratamientos y del tipo de sistemas agénticos empleados.

[39] Por ejemplo, el *Model AI governance framework for agentic AI* de la IMDA de Singapur

[40] Solo por brechas de datos personales en España de responsables que tienen obligación de notificar a la AEPD, se realizaron en 2025 más de 200 millones de comunicaciones de brechas a los ciudadanos, lo que implica que una media de cuatro brechas de datos personales fueron comunicadas a cada ciudadano español. https://www.aepd.es/prensa-y-comunicacion/notas-de-prensa/la-aepd-recibio-en-2025-mas-2.700-notificaciones-brechas

el cumplimiento y gestionar impactos críticos y riesgos para los derechos y libertades de las personas.

- ***Elementos básicos que hay que incorporar a la gobernanza de la organización***

La gobernanza de la entidad, y los procesos de gestión desarrollados a partir de ella, han de tener en cuenta las siguientes cuestiones con relación a la inclusión de sistemas de IA agéntica en los tratamientos de datos personales:

- Asignar e identificar y, en su caso integrar en los roles ya definidos en la organización, aquello roles con relación a los sistemas de IA agéntica (como responsables funcionales o responsables de IA).
- Anticipar los posibles efectos colaterales de la inclusión de IA agéntica en los tratamientos.
- Incluir las cuestiones de cumplimiento, impacto crítico y riesgo que puede implicar incluir sistemas de IA agéntica en tratamientos.
- Determinar los casos de uso para cada tratamiento y los distintos perfiles de usuarios.
- Criterios para la selección de los agentes, sus componentes y conexiones con el exterior.
- Controlar del rediseño de los tratamientos de datos personales cuando se incluyen sistemas de IA agéntica.
- Contemplar la supervisión humana cuando se necesaria.
- Realizar la necesaria adaptación de los servicios internos a los que estará conectado.
- Formalizar las relaciones con los terceros que permiten desplegar agentes (desarrolladores de modelos, proveedores de IA agéntica y otros servicios externos) asegurándose existan medidas para cumplir con sus propias responsabilidades. En particular, aclarando la distribución de obligaciones en los términos y condiciones o contratos entre la organización, los niveles de calidad de servicio y las funcionalidades para mantener el control, la privacidad, seguridad (“safety”), ciberseguridad y el control.
- Controlar el despliegue, su monitorización continua, mantenimiento y retirada de los sistemas de IA agéntica.
- Identificar los roles con respecto a protección de datos de las entidades externas.
- Prever las nuevas casuísticas que se pueden generar con relación a los derechos de acceso, rectificación, supresión, limitación del tratamiento, portabilidad y oposición, y la reacción en tiempo y forma ante dichos derechos.
- Realizar la integración con los procesos de gestión de incidentes y cumplimiento de obligaciones con relación a brechas de datos personales.
- Adaptar los planes de formación.
- Mantener una monitorización, supervisión y auditoría continua sobre los tratamientos que incorporan los servicios de IA agéntica con procedimientos de

respuesta y responsabilidad claras para actuar ante desviaciones, incidentes o incumplimientos normativos.

- Canales de información ágiles sobre actualizaciones de los sistemas IA agénticas, del uso en los tratamientos, de alternas y de incidencias que impliquen a los roles de gobernanza y, en la medida de lo necesario, a las personas usuarias.

Finalmente, adaptar o implementar políticas, y otras medidas, que se pueden enmarcar en la ejecución de los objetivos de gobernanza (ver el resto de presente capítulo).

B. EVALUACIÓN CONTINUA DEL AGENTE BASADA EN EVIDENCIAS

En la medida que se automatizan las operaciones de tratamiento, de igual forma o incluso más se tienen que automatizar la supervisión de esos procesos con respecto al cumplimiento de políticas y medidas adoptadas.

En la automatización de la auditoría se deberían incluir todas las medidas que hayan sido seleccionadas por la organización con relación a la gestión para el cumplimiento de la normativa de protección de datos. Este proceso requiere una aproximación estructurada y abarca tanto los sistemas de IA agéntica en su conjunto y en el marco del tratamiento, como la evaluación individual de cada uno de los componentes y servicios que la integran.

Esto podría incluir la revisión de funcionalidades de cada componente ante cambios en dichos elementos, con fines de cumplimiento normativo y mitigación de riesgos. Los métodos de evaluación pueden incluir pruebas de referencia (*benchmark testing*), evaluaciones con intervención humana (*human-in-the-loop*), pruebas A/B[41] y simulaciones en entornos reales.

Un aspecto crítico de esta evaluación es el conocimiento y análisis del historial de brechas e incidentes de seguridad ocurridos en los servicios evaluados y en los sistemas de IA agéntica que los incorporan.

- ***Establecimiento de criterios y métricas claras de funcionamiento***

Los criterios de funcionalidad han de permitir identificar cuándo se está comportando el sistema de IA agéntica y sus componentes de forma correcta e incorrecta y medidas objetivas que puedan servir de patrones de referencia. En particular, criterios y métricas de transparencia, reproducibilidad, control, cumplimiento y trazabilidad.

- ***Prácticas de "Golden testing"***

Consiste en disponer de un conjunto de procedimientos y datos diseñados, repetibles y preparados para comparar el resultado actual de un sistema con un resultado de referencia considerado correcto. El resultado es denominado *golden result* o *golden sample* y su aplicación forma parte de las técnicas de test de validación.

[41] Método experimental para comparar dos versiones de un elemento (como un servicio) y determinar cuál funciona mejor según métricas específicas, dividiendo aleatoriamente a los usuarios en dos grupos: uno ve la versión A (control, original) y otro la B (variante con un cambio).

Permite la comprobación repetible y la evaluación de desviaciones ante cambios en los términos legales y la funcionalidad de los sistemas, además de aumentar la explicabilidad y la transparencia del sistema en contextos bien definidos.

- ***Contratos y otros vínculos legales***

La realidad de la contratación de los servicios en Internet es que, en muchas ocasiones, los contratos no se ajustan al régimen jurídico local, los términos de los contratos cambian unilateralmente, e incluso el objeto del contrato se altera sin previo aviso (cambios de versiones, discontinuidad de funcionalidades, etc.). Además, la naturaleza dinámica de cualquier servicio y aplicación digital exige una revisión cada vez que haya una actualización de los términos y contratos, así como de aspectos técnicos de los propios servicios para determinar cómo cumplir con el RGPD.

Por ello, para aquellos componentes o servicios que tengan impacto en protección de datos, el responsable del tratamiento tiene que evaluar las condiciones tanto en el momento de las decisiones de diseño, como de forma dinámica o automática durante el ciclo de vida, para determinar los cambios legales de los componentes y una evaluación de su adecuación.

A partir de ahí, tomar las decisiones sobre cómo realizar la implementación de cada tipo de IA agéntica y para cada tratamiento.

- ***Aplicar el principio de precaución***

En el despliegue de soluciones de IA agéntica se pueden adoptar un «enfoque incremental», por ejemplo, incorporando poco a poco tratamientos, comenzando por aquellos de menor riesgo, en casos limitados, etc. También es posible ir optando por el uso de aquellos sistemas agénticos de IA ya experimentados en organizaciones similares y consultar los incidentes y problemas que ya se han originado en el entorno de la organización para poder preverlos y adoptar las medidas oportunas.

El principio de precaución también se puede aplicar a nivel de la operación de la IA agéntica, como la activación del modo de observación, en la que los agentes "observan" cómo los usuarios interactúan antes de adaptar respuestas.

- ***Explicabilidad***

En la medida que la automatización implica el uso de modelos de lenguaje es necesario realizar auditorías específicas de explicabilidad, tanto del modelo que se utiliza, como del funcionamiento conjunto de los agentes o la IA agéntica.

La explicabilidad se puede conseguir por análisis de "caja blanca" (análisis de código de los orquestadores, comprobación de flujos de datos, etc.) y mediante pruebas de "caja negra" por lo que está muy relacionado con los "*Golden test*".

- ***Intervención humana***

En el caso de que la supervisión humana tenga impacto en el tratamiento, será necesario implementar una auditoría regular sobre la eficacia de dicha supervisión.

C. MINIMIZACIÓN DE DATOS

El principio de minimización busca limitar el tratamiento de datos personales a lo estrictamente necesario, y es posible realizarlo cuando los agentes se diseñen y configuren adecuadamente para que por defecto no intenten ser eficaces simplemente mediante "fuerza de bruta" de volumen de datos.

- ***Definición de políticas de acceso a la información de la organización***

Para cada tratamiento en el que se va a utilizar la IA agéntica debe estar claramente definido que servicios y repositorios de datos pueden ser accedidos por los agentes y deben garantizarse la eficacia de tales restricciones de acceso. Es decir, implementar una política de información que incorpore el principio "*need to know*" a la IA agéntica.

Dichas políticas serán la base para la aplicación del principio de minimización (ver el apartado de Minimización) y de la gestión de la memoria interna de la IA agéntica (ver Memoria).

- ***Catálogo y catalogación de datos***

Para poder controlar la información que se dispone es necesario conocer que datos están disponibles. Conocer significa asignar una identificación que permita singularizarlos permitiendo gestionar y limitar información, por ejemplo, por etiquetas. Al singularizarlo se puede determinar cuáles son aptos o apropiados para extraerles valor en un contexto determinado de forma eficiente. Cuando se abordaron los aspectos de la memoria, ya se trató la importancia de añadir metadatos a la memoria con propósito de eficiencia en la inferencia y actuación de agentes, aunque dichos metadatos pueden también tener un papel importante como medida de protección de datos personales.

Esa identificación podrá ser a nivel de conjuntos de datos (por ejemplo, expedientes, correos) o a nivel de campos de datos (por ejemplo, destinatarios de un correo). La identificación se realiza añadiendo datos (metadatos) a los datos originales.

Por tanto, catalogación se define como un método sistemático para inventariar, organizar y gestionar los activos de datos mediante metadatos, facilitando su descubrimiento, gobernanza y uso eficiente.

Para ello, es necesario que en la catalogación se caracterice la calidad de la información almacenada (exactitud, relevancia, antigüedad, ámbito, sesgos, condicionantes normativos de uso, contexto objetivo, etc.).

Un catálogo de datos actúa como un repositorio centralizado que indexa metadatos de bases de datos, archivos y fuentes diversas, incluyendo origen, formato, propietario y linaje.

- ***Catalogación de fuentes no estructuradas***

Las fuentes no estructuradas representan un alto porcentaje de los datos en una entidad (por ejemplo, correos electrónicos, actas y grabaciones de reuniones, informes, etc.), y se caracterizan por carecer de formato fijo, complicando su indexación, escalabilidad y búsqueda, además de requerir altos recursos para volumen masivo.

Estrategias para la catalogación de datos no estructurados pasan por el enriquecimiento con metadatos, etiquetado automatizado o estructuración de datos no estructurados. Para ello, se emplean técnicas basadas en NLP, análisis de audio y video, búsqueda semántica de patrones, recuperación contextual, herramientas DPL (*data loss prevention*) para identificar y clasificar fuentes de información que incorporan datos personales (y sensibles o confidenciales), etc.

A partir de ahí, se podría realizar un preproceso de la información para la extracción de datos específicos para las tareas de los agentes. En particular, de anonimización o retirada de datos personales que no sean necesarios.

- ***Granularidad de la minimización***

La minimización tiene como objeto tratar los datos que solo sean necesarios en el tratamiento. La aplicación de dicho principio tiene dos niveles de granularidad, a nivel de tratamiento y a nivel de operaciones de tratamiento.

Por ejemplo, en el marco, de la agéntica, en un tratamiento de respuesta automática a mensajes de correo electrónico, aplicar minimización en la operación revisión sintáctica de un mensaje de correo electrónico antes de ser enviado implica no tratar el nombre del destinatario, y en la operación enviar el correo implica tratar dicho nombre, pero no analizar el contenido del mensaje[42].

La minimización debe enfocarse tanto a datos de sujetos en general, como a la información de las mismas personas usuarias de los sistemas. En particular:

- Un diseño que evite el perfilado personal del usuario
- Eliminación de metadatos que no son útiles en el pipeline o por fases del *pipeline*.
- Desvinculación de acciones del usuario cuando no sean necesarias.

- ***Filtrado de flujos de datos***

Con relación a lo anteriormente expuesto, el análisis no solo se puede hacer en los datos en reposo, sino se podría plantear con los datos en tránsito entre las distintas acciones del agente cuando impliquen comunicación de datos con terceros y externos. Es decir, en las etapas intermedias de las cadenas de razonamiento se podría incorporar filtrado de la información intercambiada para la categorización, higienización, minimización y alerta en los intercambios de datos.

Esto no solo evitaría detectar, por ejemplo, inyecciones de *prompt* que se generen internamente, sino también exposición de datos personales, uso excesivo de datos, accesos masivos a información, etc. Otro caso, para determinar si en la operación de los agentes se está produciendo el tratamiento de categorías especiales de datos que no sean necesarias para el tratamiento.

[42] En tratamiento de información clasificada se conoce como el principio de “need-to-know” o necesidad de conocer que tienen cada uno de los intervinientes en un tratamiento.

En estos caos, la utilización de inteligencia artificial, por ejemplo utilizando modelos pequeños de lenguaje (SML), en conjunto con patrones de detección de amenazas a la protección de datos y a la seguridad podrían ser técnicas aplicables.

- ***Shadow leaks***

Con el objeto de minimizar el "*shadow leaks*"[43] se deberán implementar medidas, como el uso de herramientas DPL (*data loss prevention*) orientadas a minimizar la exposición del contexto interno del sistema, limitar la divulgación de explicaciones relativas a razonamientos o reglas de funcionamiento, aplicar controles de correlación entre consultas para detectar relaciones indebidas, emplear respuestas genéricas frente a preguntas de carácter meta[44] y monitorizar de forma continua los patrones de consulta a largo plazo con el fin de identificar comportamientos anómalos o riesgos potenciales.

- ***Seudonimización de las personas usuarias***

Seudonimizar la interacción del usuario con el agente, de modo que se utilicen *tokens* de un solo uso para la interacción entre los componentes o para los accesos a servicios externos cuando se requiera autenticación. Esto permitirá, entre otros, evitar el control y perfilado de las mismas (ver siguiente apartado) además que evitará la efectividad de que el agente de IA otorgue consentimientos efectivos a servicios externos, cree nuevas relaciones entre usuario y otros responsables o firme contratos.

- ***Control y perfilado de las personas usuarias***

La memoria de la IA agéntica, así como los ficheros de registro de los diversos componentes y servicios utilizados por las personas usuarias (p.ej. empleados del responsable), pueden recopilar y almacenar información, que puede incluso suponer un perfil de los mismos. Para ello se podría considerar:

- Tener una política de recogida de información de la interacción del usuario con el agente en la memoria de corto y largo plazo limitada a los aspectos relevantes para cada tratamiento específico.
- Recoger en los ficheros de registro la información imprescindible para un nivel de trazabilidad y seguridad adecuados.
- Seudonimizar la información de registro.
- Seudonimizar la interacción del usuario con el agente como se ha descrito en el apartado anterior.
- Plazos de caducidad para la información recogida en los ficheros de registro y en históricos de memoria a largo plazo.

[43] Son situaciones donde datos sensibles, patrones internos o información confidencial se infieren o se exponen sin que exista una "fuga" explícita, por ejemplo, a través de metadatos, tiempos de respuesta o comportamiento del sistema, mediante salidas parciales que permiten reconstruir información privada y otros. En general, un *shadow leak* no es una filtración directa, sino una exposición silenciosa y difícil de detectar que surge como efecto secundario del funcionamiento normal del sistema.

[44] Aquellas consultas que no buscan directamente la información funcional del sistema, sino que intentan obtener conocimiento sobre su funcionamiento interno, sus reglas, sus fuentes, sus mecanismos de razonamiento, sus límites o sus salvaguardas. Este tipo de preguntas operan a un nivel "meta" porque analizan o explotan el propio sistema en lugar del dominio de información que este ofrece.

D. CONTROL DE LA MEMORIA

El control de la memoria del sistema de IA agéntica está muy relacionado con las estrategias de minimización de datos, las garantías de explicabilidad y repetitividad de inferencias o perfilados de personas y la capacidad de trazabilidad para aplicar gestión del consentimiento, ejercicios de derechos y limitación del tratamiento.

El control de la memoria del agente ha de realizarse tanto sobre la memoria a corto plazo como sobre la memoria a largo plazo.

- ***Gestión de memoria***

Introducir la capacidad de acceder, tener catalogado y gestionar el contenido de la memoria permitiendo, por ejemplo, búsqueda por contenido y parámetros de calidad, borrado, establecer limitaciones de tratamiento o alertas de uso, incluir trazabilidad de los accesos, auditable, etc.

- ***Compartimentación de la memoria***

En el caso de una misma IA agéntica en la organización, hay que contemplar la oportunidad de tener la memoria compartimentada y gestionada para distintos tratamientos, distintos casos dentro de los tratamientos y/o para distintas personas usuarias.

El nivel de granularidad de la compartimentación dependerá de los tratamientos, definiendo claramente que memoria será de uso común a cualquier operación de la IA agéntica en la organización ya que implementa políticas de esta, y que datos e información será necesario que esté separada entre los tratamientos, los usuarios y los distintos casos. La rigidez de dicha compartimentación, desde un división física, un división lógica rígida o una búsqueda por catalogación dependerá del tratamiento y la política del responsable.

- ***Análisis y filtrado de la memoria de la persona usuaria***

Es necesario poder limitar los efectos que la memoria de la persona usuaria pueda tener en aspectos sustanciales del tratamiento, aspectos que ya han sido identificado el responsable. Para ello, es necesario separar aspectos de personalización en la ejecución de las tareas de aspectos que puedan tener incidencia en la aplicación de políticas de la organización, coherencia entre distintas actuaciones de la organización o aparición de sesgos.

Para ello es necesario poder diferenciar entre la memoria de la organización, gestionada por los servicios TIC, y la memoria de la persona usuaria para que esta última no sea tenida en cuenta en ciertas acciones que pueda realizar la IA agéntica. Estas limitaciones dependerán de cada tratamiento y podrían ser, por ejemplo, sobre la división en subtareas, acceso a determinadas herramientas o sobre decisiones finales.

- ***No log policy selectivo***

Cuando un sistema de IA agéntica se utiliza para implementar distintos tratamientos con alguno de sus componentes, por ejemplo los LLMs, implementando registros o logs

donde almacenarán la actividad de todos los tratamientos, es aconsejable utilizar una política de "no log" o política de cero retención de datos a nivel de componente.

Dicha política supone que el registro de información en el componente es mínimo, y únicamente relacionado con el origen de las peticiones y el tipo, pero no su contenido. Por ejemplo, el componente de inferencia no almacenaría el contenido de los prompts o las inferencias, que sí pueden estar registradas a nivel de log del tratamiento, para cada tratamiento de forma independiente, y conforme a las políticas de información del responsable.

- ***Establecimiento de plazos de retención estrictos***

Fijar plazos y establecer procedimientos para la eliminación de datos por categorías específicas y diferenciados según las necesidades de cada uno de los componentes que conforman el tratamiento utilizando IA agéntica.

- ***Desactivación del almacenamiento en memoria***

En determinados tratamientos y según sus necesidades, permitir la desactivación de la memoria persistente por defecto o su desactivación por la persona usuaria[45]. La granularidad de la desactivación podrá ser a nivel de subtareas que se puedan considerar de alto riesgo para evitar el almacenamiento de datos personales irrelevantes para futuros tratamientos o evitar la persistencia de inyecciones maliciosas.

- ***Aplicar estrategias de higienización de la memoria***

Aplicar técnicas de higienización o depuración[46] de la memoria a largo plazo mediante comprobación automática de contenido dañino, caducidad de entradas sin uso u obsoletas, análisis de la coherencia de la información, búsqueda y eliminación de credenciales del usuario innecesarias, destilado de la información, análisis y eliminación de sesgos además de estrategias para obligar al usuario/administrador a realizar limpiezas periódicas.

E. AUTOMATIZACIÓN

- ***Decisión sobre el grado de autonomía***

El grado de autonomía que puede tener el sistema IA agéntico deberá ser establecido por el responsable para cada uno de los tratamientos teniendo en cuenta el contexto, el ámbito, las finalidades y el riesgo que puede suponer para los derechos y libertades de las personas, y el cumplimiento normativa con relación a las decisiones automatizadas, debiendo estar apropiadamente justificado basado en evidencias y documentada la decisión.

[45] Esto es distinto a enviar un *prompt* diciendo que la información que sí ha sido almacenada no se tenga en cuenta.
[46] En inglés el concepto se define como "*sanitization*".

- ***Diseño eficaz y seguro de las cadenas de razonamiento***

El diseño de las cadenas de razonamiento deberá estar controlado y validado. En el caso de que la cadena de razonamiento se elabore mediante LLMs, hay que evaluar la capacidad que tienen el nivel de calidad necesario para abordar los contextos de los tratamientos en los que se va a emplear la IA agéntica. Además, hay que garantizar que en la elaboración de la cadena de razonamiento no hay posibilidad de contaminación entre distintos modelos aprendidos no compatibles (por ejemplo, subtareas de procedimientos administrativos de distintas jurisdicciones).

En su caso, evaluar la necesidad de implementar las cadenas de razonamiento, de forma total o un nivel superior de abstracción, de forma codificada (*hardcoded)* por el administrador. Por ejemplo, hacer una división de un tratamiento en subtareas[47] manualmente, y dejar que los agentes de razonamiento elaboren el detalle de dichas subtareas.

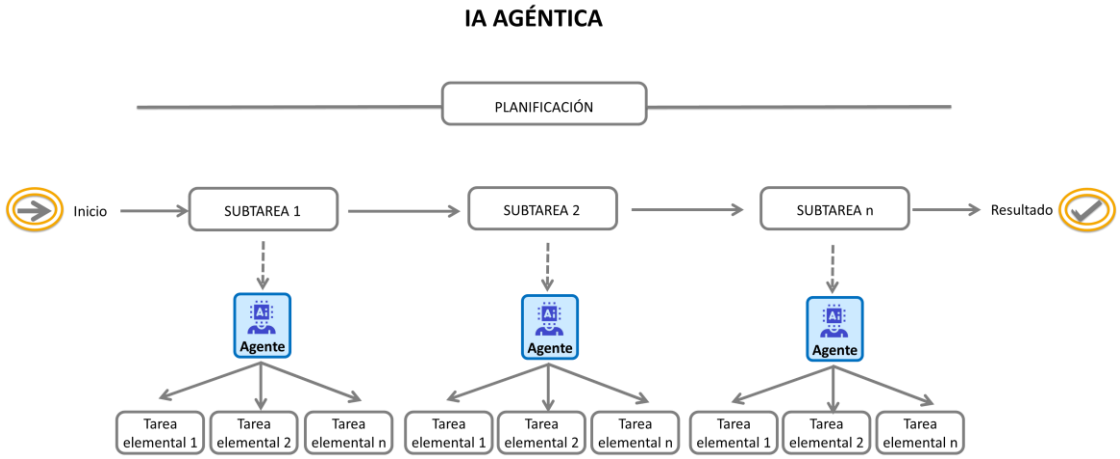

Figura 14 Ejemplo de dos niveles de descomposición de tareas

Es necesario prever la posible aparición de ataques por inyección de *prompts* y la generación de errores compuestos. Entre otros, se deberían establecer controles que garanticen la separación estricta entre datos e instrucciones, la correcta etiquetación y trazabilidad del origen del contenido, la limitación de privilegios de las herramientas utilizadas, la validación y sanitización de las entradas en cada etapa del proceso (en particular de la información en la memoria persistente) mediante *guardrails*[48].

También se podría realizar una evaluación automática de las decisiones finales adoptadas por el agente, e incluyendo las decisiones parciales en puntos críticos susceptibles que generar errores compuestos y la inclusión de mecanismos de evaluación de la confianza en la memoria persistente.

[47] Por ejemplo, fijar en un despacho de abogados los pasos de un proceso judicial, y dejar a los agentes el razonamiento de la resolución de cada paso.

[48] Mecanismos de seguridad y restricciones que guían el comportamiento de los modelos de IA para evitar salidas dañinas, sesgadas o inapropiadas, similar a las barreras en una carretera que impiden desviarse.

- ***Catálogo y listas blancas de servicios***

Se trata de disponer de un catálogo de servicios, que pueden incluir distintos LLMs, en el que se identifiquen versiones y, en particular, la fiabilidad que se da a cada servicio y/o adecuación para distintos contextos, así como evitar el efecto de alucinaciones que realicen llamadas a servicios inexistentes.

Esto permite que, para distintos contextos, se utilice como lista blanca, con flexibilidad para el uso de sistemas de IA agéntica en distintos tratamientos (para los desarrolladores, en caso de que las llamadas a funciones esté predefinida, o para limitar las herramientas invocables por el LLM en caso contrario). El catálogo debería abarcar servicios externos pero también internos, por ejemplo, repositorios de datos o servicios que permiten el acceso a la pantalla del operador.

- ***Limitación de servicios accesibles***

La limitación de servicios accesibles podría complementar al catálogo anterior por tratamientos específicos. De esta forma, cada tratamiento tendría políticas definidas sobre el máximo tipo de herramientas y accesos a datos que necesitaría para completar las tareas. Por ejemplo, en operaciones de consulta de normativa no se requeriría el acceso a servicios web si se dispone a priori de una recopilación actualizada disponible en forma de RAG.

- ***Control en la ejecución de herramientas***

La invocación de herramientas y servicios en Internet son *de facto* salidas parciales de la IA agéntica que pueden ser transparentes para las personas usuarias. Los controles que se podrían establecer sobre ellas son:

- Control de los parámetros con los que se invocan las herramientas, implementando *guardrails* y formatos rígidos para detectar parámetros erróneos o sesgados.
- Control de la respuesta de las herramientas, implementando nuevos *guardrails* sobre el contenido de entrada.
- Obligación de una acción humana de supervisión con relación a determinadas herramientas que pudieran tener un mayor impacto.

- ***Criterios y puntos de control para la intervención humana***

Desde el diseño se deberían definir criterios y puntos de control significativos o límites de actuación que requieran aprobación humana, especialmente antes de que se ejecuten acciones sensibles. Esto puede incluir:

- Acciones y decisiones de alto impacto, por ejemplo, la edición de datos sensibles, decisiones finales en ámbitos de alto riesgo (como la atención sanitaria o el ámbito jurídico), o acciones que puedan generar responsabilidad legal. Otro ejemplo podría ser la utilización de credenciales de usuario obtenidas de la memoria, solicitar una autorización previamente a su uso en subtareas críticas.
- Acciones irreversibles, por ejemplo, la eliminación permanente de datos, el envío de comunicaciones o la realización de pagos.

- Comportamientos atípicos o fuera de lo común, por ejemplo, cuando un agente accede a un sistema o base de datos fuera de su ámbito de trabajo, o cuando selecciona una ruta de entrega que duplica la distancia mediana.
- Definidas por el usuario. Los agentes pueden actuar en nombre de usuarios con diferentes niveles de tolerancia al riesgo. Además de los límites definidos por la organización, puede ofrecerse a los usuarios la opción de establecer sus propios límites, por ejemplo, exigir aprobación para compras que superen una determinada cantidad.

- ***Reversibilidad de las acciones de los agentes de IA***

Evaluar la necesidad de implementar medidas que permitan revertir determinadas acciones, por ejemplo si el agente puede modificar datos personales.

- ***Nivel de autonomía de acuerdo al tratamiento***

Incluir controles de la capacidad de autonomía ajustables para cada tratamiento que esté utilizando la IA agéntica, en función del impacto o por riesgo, desde ejecución autónoma en tratamientos de impacto y riesgo bajo hasta intervención humana obligatoria con gran granularidad en las operaciones del tratamiento cuando sean altos.

- ***Supervision humana efectiva***

Como es necesario, es necesario determinar cuándo integrar a expertos en puntos críticos del flujo para validar, refinar o anular las decisiones del agente (con los mecanismos de anulación oportunos) antes de que tengan un impacto real.

Para la evaluación de la intervención humana se aconseja tener en cuenta:

- Competencia y autoridad: tiene la autoridad o la tarea asignada que le permite alterar el resultado de la decisión automatizada.
- Preparación y formación: tiene la capacidad y las aptitudes para evaluar la decisión y los factores que determinan esa decisión con relación al contexto del tratamiento y el sistema automatizado empleado, en sus capacidades y limitaciones.
- Independencia: evaluar si existen presiones desde la organización o desde fuera de la organización que condicionen la disputa de la decisión por parte de la persona.
- Diligencia en el ejercicio de su competencia: en particular, si está sometido al sesgo de automatización.
- Medios para poder ejercer su competencia y cualificación.
    - Que existan los procedimientos y los medios técnicos para intervenir, en el momento o punto adecuado en tiempo y forma.
    - Que disponga de la información necesaria en tiempo y forma para poder ejercer su cualificación, en particular, conocer las consecuencias, riesgos de las decisiones en general, y las que se están tomando para los casos específicos y todos los aspectos que condicionan la decisión automatizada.

Estos incluyen los datos del individuo concreto, pero también podrían incluir los procedimientos para la recogida de datos de entrada, los datos implícitos en el modelo que genera la decisión, los datos contextuales que no se han tenido en cuenta en la decisión automatizada, además de las capacidades y límites del sistema de decisión. También aquellos datos que la persona, en su cualificación, estime que son necesarios contemplar para el caso concreto y que no se han considerado en la decisión automatizada.

- Que disponga de los recursos para poder ejercer su cualificación: las decisiones de la IA agéntica deben ser explicables por ejemplo, aplicaciones que le permitan analizar la información en el formato que se esté utilizando para la decisión automatizada, etc.
- Que disponga del tiempo necesario para poder ejercer su cualificación para cada una de las decisiones que sean de su competencia.

- ***Rutas de escalamiento***

La supervisión humana podría complementarse con una monitorización automatizada en tiempo real para escalar cualquier comportamiento inesperado o anómalo. El escalamiento supone la implementación de protocolos y técnicas para transferir el control de procesos automatizados a un operador humano cuando se detectan situaciones de alto riesgo, incertidumbre o anomalías.

Esto puede lograrse mediante la implementación de alertas para determinados eventos registrados (por ejemplo, intentos de acceso no autorizado a datos personales, o múltiples intentos fallidos de invocar una herramienta), el uso de técnicas de ciencia de datos para identificar trayectorias anómalas de los agentes, el uso de agentes para supervisar a otros agentes, el acceso a categorías especiales de datos cuando no es necesario, etc.

- ***Principio de los cuatro ojos***

En casos de procesos automáticos con gran impacto en los derechos y libertades de las personas, se puede plantear aplicar el principio de doble verificación por distintas personas, que constituyen una capa adicionales de confianza en el mecanismo de supervisión humana y fomentan la conciencia crítica del operador.

F. CONTROL DEL AGENTE DESDE EL DISEÑO

El agente de IA va a permitir automatizar todo o en parte un tratamiento, por lo tanto, podría ser necesario rediseñar el tratamiento para desplegar el sistema IA dentro de él con garantías. En este apartado, se enumeran medidas de control del agente que se podrían incluir en el tratamiento, y que la IA agéntica seleccionada debería permitir implementar.

- ***Documentación***

Mantener un registro con control de integridad (no necesariamente un papel) del proceso de responsabilidades, decisiones, acciones tomadas, diseños, arquitectura, eventos de explotación y evolución, mantenido dinámicamente.

- ***Profesionales cualificados***

Utilizar para el despliegue de sistemas de IA agéntica en el tratamiento un equipo de profesionales cualificados; no solo se trataría de poner en marcha un agente de IA, sino que hay que tener en cuenta las implicaciones que tienen la automatización de procesos organizativos, y por ello se requiere personal con conocimientos en ciencia de datos, calidad de procesos, contexto de operación, seguridad y cumplimiento normativo, entre otros.

- ***Trazabilidad***

La trazabilidad del dato es la capacidad de conocer todo el ciclo de vida del dato: la fuente del dato, la fecha y hora exacta de extracción, cuándo, dónde y por quién se produjo su transformación, y cuándo, dónde, por quién y con qué finalidad y legitimidad se cargó en un repositorio, se usó o se descargó desde un entorno a otro repositorio. A este proceso también se conoce como "*Data Linage*".

En este sentido, cuanto más complejo es el ciclo de vida del dato y más intervinientes participan en él, más valor tiene incorporar trazabilidad en el tratamiento.

La trazabilidad puede cumplir propósitos ajenos a protección de datos, como control de los secretos comerciales, la propiedad intelectual e industrial, perfeccionamiento de contratos. Por otro lado, podrá cumplir con los siguientes objetivos desde el punto de vista del RGPD:

- Cumplir con los requisitos de transparencia a los interesados del RGPD.
- Permitir el ejercicio efectivo de los derechos de los interesados, en particular, la gestión del consentimiento.
- Permitir ejercer las obligaciones del responsable del tratamiento (p.ej. para que garantice los principios de limitación de tratamiento, las finalidades ajustadas a las bases jurídicas o el control de encargados/subencargados de tratamiento).
- Tener evidencia de que datos se tratan en cada operación de tratamiento ejecutada en la IA agéntica, en sus fases intermedias. En particular si se están utilizando categorías especiales de datos.
- Controles sobre los empleados que participan en el tratamiento, ahora como usuarios de la IA agéntica, para prevenir abusos y sesgos.
- Demostrar diligencia y transparencia a los sujetos de los datos y a las autoridades de Control.

Por lo tanto, las medidas para garantizar dichas capacidades están relacionadas con la catalogación de datos, y supone guardar registros (*logs*) de la información procesada por todos los procesos de razonamiento, las fuentes accedidas y los servicios empleados

en la inferencia, tanto de entrada como de salida. En particular, poder tener un control detallado de datos y propósitos para los que servicios externos acceden a información.

Esto es especialmente relevante tanto para transparencia en el tratamiento de datos, como con propósitos de análisis de la reproducibilidad de las inferencias, el control de la información que de la persona usuaria es tratada por servicios, para control de cumplimiento normativo, poder implementar políticas de información, etc.

- ***Test de verificación y validación***

Aunque los test de verificación y validación son técnicas muy conocidas en ingeniería de sistemas, y no específicas de sistemas de inteligencia artificial, se considera importante recordar que existen y siguen siendo aplicables en el despliegue de sistemas de IA agéntica. Son una herramienta clave para implementar transparencia en la cadena de valor, explicabilidad y garantizar la robustez.

La verificación consiste en comprobar si el sistema se está construyendo correctamente, es decir, si cumple con los requisitos, especificaciones de diseño y estándares mediante técnicas estáticas como revisiones, inspecciones y análisis de código (por ejemplo, comprobando que los flujos de datos internos y externos son realmente los declarados). La validación comprueba que las necesidades reales del usuario en un contexto determinado se cumplen con relación a las métricas de calidad establecidas, mediante pruebas dinámicas con ejecución del código, como pruebas funcionales, de integración y de aceptación.

- ***Definir y controlar que los prompts siguen un procedimiento operativo estándar***

Definir un procedimiento operativo estándar (SOP de *Standard Operating Procedure*) para construcción de *prompts*. Esto supone definir un conjunto estructurado de instrucciones paso a paso que detallan cómo debe actuar un agente de IA en el marco del tratamiento para lograr resultados consistentes y más predecibles y evitar prompt maliciosos.

Por ejemplo, se podría definir que los *prompt* se estructurasen de la siguiente forma: interpretación inicial, clasificación, criterio de validación, descomposición preliminar del problema, selección de herramientas, criterios de búsqueda de información, de verificación cruzada, de limpieza de datos, de evaluación, etc. Todo ello con campos predefinidos.

La aplicación de SOP mediante *front-ends* con campos validados adaptados a cada tratamiento permite un uso eficaz de esta medida. En cualquier caso, no desplaza en todos los tratamientos el uso de medidas de control de la memoria y de la automatización.

- ***Mecanismos de repetibilidad***

Establecer mecanismos que permitan la repetibilidad de una decisión. Por ejemplo, manteniendo un registro de la configuración en un proceso de decisión: las entradas de datos que han generado una decisión final, el tráfico intermedio de datos en la cadena

de razonamiento, así como las configuraciones de seudoaleatoriedad en los sistemas "probabilistas" y otros valores[49].

A su vez, poder reintroducir dichos valores en la IA agéntica y realizar pruebas de funcionamiento, lo que repercute en la transparencia y explicabilidad del agente.

- ***Gestión de identidad, autenticación, y privilegios***

La gestión de identidad digital de usuarios, de la IA agéntica y de sus componentes es una herramienta de trazabilidad y auditoría. Además, permite la gestión necesaria para prevenir la escalada no autorizada de privilegios de los agentes, la suplantación de identidad y las violaciones de control de acceso.

El principio básico a aplicar en el entorno de la IA agéntica es el de menor privilegio, y aplicar las siguientes estrategias:

- Implementar mecanismos seguros de autenticación tanto para las personas usuarias como para la IA agéntica y sus componentes. : por ejemplo, requerir verificación criptográfica de identidad para los agentes, implementar RBAC y ABAC granulares, autenticación multifactor (MFA) para cuentas con altos privilegios, forzar la re-autenticación continua en largas sesiones, evitar la delegación de privilegios entre agentes excepto autorizada en flujos predefinidos, autenticación mutua en interacciones IA-a-IA y entre agentes, limitar a persistencia de credenciales o temporalidad de las credenciales de los agentes, etc.
- Restringir la escalada de privilegios y la herencia de identidad: por ejemplo utilizar controles de acceso dinámicos que expiren los permisos elevados, elaborar perfiles de comportamiento basados en IA para detectar inconsistencias en la asignación de roles y en los patrones de acceso de los agentes, exigir validación por un humano para acciones de IA de alto riesgo que impliquen cambios en la autenticación, detectar car anomalías de herencia de roles en tiempo real, aplicar restricciones temporales a la elevación de privilegios, etc.
- Detectar y bloquear intentos de suplantación de IA: como por ejemplo, detectar inconsistencias en la verificación de identidad, supervisar cambios inesperados de rol, detectar, registrar y alertar de desviaciones sospechosas en los intentos de autenticación o intentos fallidos, así como patrones de ejecución en cascada o recursivos de herramientas activados entre agentes, aislar a los agentes que generen tráfico de protocolo sospechoso.

- ***Control estricto sobre las actualizaciones.***

Tener bajo control sobre qué actualizaciones se producen en cada elemento del sistema de IA agéntica y poder de decisión sobre cuando esas actualizaciones entran en producción para evitar incompatibilidades, inestabilidades y falta de robustez, nuevos tratamientos, cambios en la funcionalidad, aparición de nuevas vulnerabilidades, cambios legales con incumplimiento normativo, transferencias internacionales no controladas, etc.

[49] Como semillas, parámetros de temperatura, Relevancia Marginal Máxima (MMR) en RAGs, etc.

Un mecanismo de prevención es el incluir sistemas de control de versiones con posibilidad de realizar “roll-back”.

Tener en cuenta, en caso de actualizaciones, la utilización de *sandboxing* (ver apartado más adelante) y evaluación continua (apartados anteriores).

- ***Sandboxing[50] en desarrollo y explotación***

Con relación a la capacidad de percepción y acción sobre el entorno, se pueden utilizar medidas que restrinjan la amplitud del contexto exterior con el que interacciona el agente.

En su expresión más restrictiva, la aplicación del principio del *sandboxing* implicaría la implementación de entornos de tratamiento seguro (*Secure Processing Environtments* o SPE[51]). En su expresión más laxa nos encontraríamos con una implementación sin restricciones en los permisos del sistema de IA agéntica para interactuar con el entorno. El empleo de sandboxing en la ejecución de herramientas invocadas por la IA agéntica es una aplicación intermedia común. En función de las obligaciones de cumplimiento, impactos o riesgos se deberá establecer una arquitectura entre ambos extremos.

Una posible implementación de *sandboxing* es el utilizar entornos confinados, como contenedores o microVMs, para el aislamiento en la ejecución de agentes. También el uso de técnicas de terminal restringido: entornos controlados donde el conjunto de comandos, servicios y acceso a la red está limitado a operaciones previamente autorizadas.

Este tipo de entornos son imprescindibles en las fases de prueba de despliegue.

- ***Protocolos de detección de errores y planes de contingencia***

Inclusión en la gestión de los tratamientos de procedimientos en los que se detalle qué acciones, y por quién, además de los recursos necesarios para hacer frente a un problema en la IA agéntica. En otros, con relación a brechas de datos personales, la reducción del impacto y la comunicación a los afectados.

- ***Control de flujo de extracción de datos***

Introducir, en aquellos tratamientos en los que sea necesario controles que exijan acciones expresas del usuario para comunicaciones de datos a terceros o envíos masivos de datos.

---

[50] Evitar la confusión con sanboxes o entornos controlados de pruebas regulatorios, como los definidos en el Reglamento de Inteligencia Artificial

[51] El Reglamento de Gobernanza de Datos define los entornos de tratamiento seguros en el art.2.20) como “«entorno de tratamiento seguro», el entorno físico o virtual y los medios organizativos para garantizar el cumplimiento del Derecho de la Unión, como, por ejemplo, el Reglamento (UE) 2016/679, en particular, por lo que respecta a los derechos de los interesados, los derechos de propiedad intelectual y la confidencialidad comercial y estadística, la integridad y la accesibilidad, así como para garantizar el cumplimiento del Derecho nacional aplicable y permitir que la entidad encargada de proporcionar el entorno de tratamiento seguro determine y supervise todas las acciones de tratamiento, incluida la presentación, el almacenamiento, la descarga y la exportación de datos, así como el cálculo de datos derivados mediante algoritmos computacionales;2.

- ***Cortacircuitos y límites duros de pasos***

Los cortacircuitos (o *circuit breakers*) en IA agéntica son mecanismos de seguridad programados que interrumpen automáticamente la ejecución de un agente cuando detectan anomalías predefinidas, como bucles infinitos, desviaciones de objetivos, acceso masivo a datos, intento de intercambios masivos de información, desviación de objetivos, etc.

- ***Controles de calibrado y alineación***

Los problemas de calibrado, en la medida que pueden afectar al tratamiento de datos personales (excesivos, sensibles, inexactos, sin legitimación) se pueden evitar introduciendo medidas entre las fases intermedias de las cadenas de razonamiento que evalúen acciones y datos con relación a parámetros de calidad, alineamiento con políticas y normativas e intereses empresariales. Estas medidas podrían implantarse en función del impacto o el riesgo que pueda suponer cada etapa, la falta de transparencia o explicabilidad del componente utilizado (por ejemplo, un LLM), u otros factores. También entre las medidas posibles de este tipo podría incluirse la supervisión humana.

G. GESTIÓN DEL CONSENTIMIENTO

En el caso de que haya tratamientos basados en el consentimiento, el sujeto de los datos también ha de tener la posibilidad de dar su consentimiento, modificarlo o retirarlo en el marco de una cadena de razonamiento compleja, que puede incluir múltiples repositorios y fuentes de datos de múltiples entidades.

H. DEPENDIENDO DE LA COMPLEJIDAD DE LOS TRATAMIENTOS, SU IMPACTO Y SUS RIESGOS, LA GESTIÓN PODRÍA ADOPTAR DISTINTAS FORMAS, MUY RELACIONADAS CON LO EXPUESTO EN LOS APARTADOS DE "VII.D. CONTROL DE LA MEMORIA

El control de la memoria del sistema de IA agéntica está muy relacionado con las estrategias de minimización de datos, las garantías de explicabilidad y repetitividad de inferencias o perfilados de personas y la capacidad de trazabilidad para aplicar gestión del consentimiento, ejercicios de derechos y limitación del tratamiento.

El control de la memoria del agente ha de realizarse tanto sobre la memoria a corto plazo como sobre la memoria a largo plazo.

- ***Gestión de memoria***

Introducir la capacidad de acceder, tener catalogado y gestionar el contenido de la memoria permitiendo, por ejemplo, búsqueda por contenido y parámetros de calidad, borrado, establecer limitaciones de tratamiento o alertas de uso, incluir trazabilidad de los accesos, auditable, etc.

- ***Compartimentación de la memoria***

En el caso de una misma IA agéntica en la organización, hay que contemplar la oportunidad de tener la memoria compartimentada y gestionada para distintos

tratamientos, distintos casos dentro de los tratamientos y/o para distintas personas usuarias.

El nivel de granularidad de la compartimentación dependerá de los tratamientos, definiendo claramente que memoria será de uso común a cualquier operación de la IA agéntica en la organización ya que implementa políticas de esta, y que datos e información será necesario que esté separada entre los tratamientos, los usuarios y los distintos casos. La rigidez de dicha compartimentación, desde un división física, un división lógica rígida o una búsqueda por catalogación dependerá del tratamiento y la política del responsable.

- ***Análisis y filtrado de la memoria de la persona usuaria***

Es necesario poder limitar los efectos que la memoria de la persona usuaria pueda tener en aspectos sustanciales del tratamiento, aspectos que ya han sido identificado el responsable. Para ello, es necesario separar aspectos de personalización en la ejecución de las tareas de aspectos que puedan tener incidencia en la aplicación de políticas de la organización, coherencia entre distintas actuaciones de la organización o aparición de sesgos.

Para ello es necesario poder diferenciar entre la memoria de la organización, gestionada por los servicios TIC, y la memoria de la persona usuaria para que esta última no sea tenida en cuenta en ciertas acciones que pueda realizar la IA agéntica. Estas limitaciones dependerán de cada tratamiento y podrían ser, por ejemplo, sobre la división en subtareas, acceso a determinadas herramientas o sobre decisiones finales.

- ***No log policy selectivo***

Cuando un sistema de IA agéntica se utiliza para implementar distintos tratamientos con alguno de sus componentes, por ejemplo los LLMs, implementando registros o logs donde almacenarán la actividad de todos los tratamientos, es aconsejable utilizar una política de “no log” o política de cero retención de datos a nivel de componente.

Dicha política supone que el registro de información en el componente es mínimo, y únicamente relacionado con el origen de las peticiones y el tipo, pero no su contenido. Por ejemplo, el componente de inferencia no almacenaría el contenido de los prompts o las inferencias, que sí pueden estar registradas a nivel de log del tratamiento, para cada tratamiento de forma independiente, y conforme a las políticas de información del responsable.

- ***Establecimiento de plazos de retención estrictos***

Fijar plazos y establecer procedimientos para la eliminación de datos por categorías específicas y diferenciados según las necesidades de cada uno de los componentes que conforman el tratamiento utilizando IA agéntica.

- ***Desactivación del almacenamiento en memoria***

En determinados tratamientos y según sus necesidades, permitir la desactivación de la memoria persistente por defecto o su desactivación por la persona usuaria. La

granularidad de la desactivación podrá ser a nivel de subtareas que se puedan considerar de alto riesgo para evitar el almacenamiento de datos personales irrelevantes para futuros tratamientos o evitar la persistencia de inyecciones maliciosas.

- ***Aplicar estrategias de higienización de la memoria***

Aplicar técnicas de higienización o depuración de la memoria a largo plazo mediante comprobación automática de contenido dañino, caducidad de entradas sin uso u obsoletas, análisis de la coherencia de la información, búsqueda y eliminación de credenciales del usuario innecesarias, destilado de la información, análisis y eliminación de sesgos además de estrategias para obligar al usuario/administrador a realizar limpiezas periódicas.

I. AUTOMATIZACIÓN

- ***Decisión sobre el grado de autonomía***

El grado de autonomía que puede tener el sistema IA agéntico deberá ser establecido por el responsable para cada uno de los tratamientos teniendo en cuenta el contexto, el ámbito, las finalidades y el riesgo que puede suponer para los derechos y libertades de las personas, y el cumplimiento normativa con relación a las decisiones automatizadas, debiendo estar apropiadamente justificado basado en evidencias y documentada la decisión.

- ***Diseño eficaz y seguro de las cadenas de razonamiento***

El diseño de las cadenas de razonamiento deberá estar controlado y validado. En el caso de que la cadena de razonamiento se elabore mediante LLMs, hay que evaluar la capacidad que tienen el nivel de calidad necesario para abordar los contextos de los tratamientos en los que se va a emplear la IA agéntica. Además, hay que garantizar que en la elaboración de la cadena de razonamiento no hay posibilidad de contaminación entre distintos modelos aprendidos no compatibles (por ejemplo, subtareas de procedimientos administrativos de distintas jurisdicciones).

En su caso, evaluar la necesidad de implementar las cadenas de razonamiento, de forma total o un nivel superior de abstracción, de forma codificada (*hardcoded)* por el administrador. Por ejemplo, hacer una división de un tratamiento en subtareas manualmente, y dejar que los agentes de razonamiento elaboren el detalle de dichas subtareas.

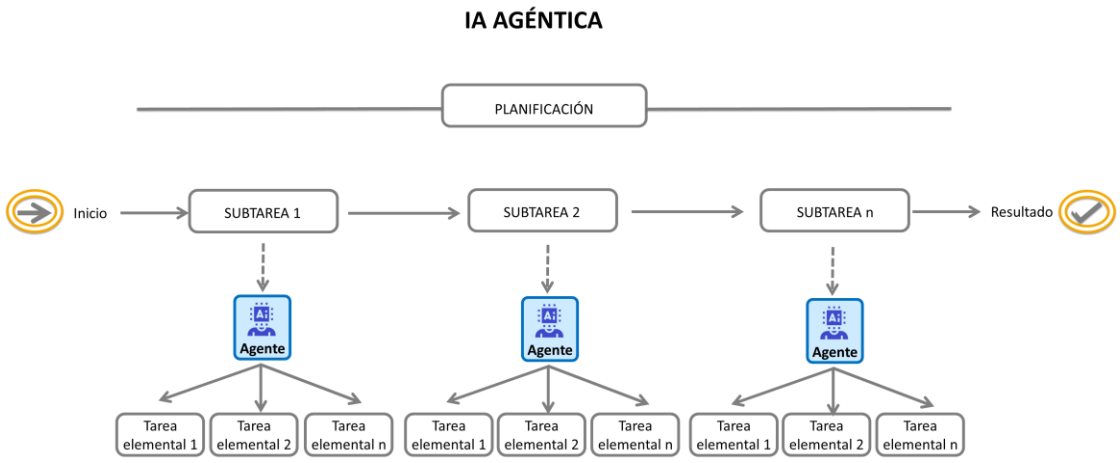

Figura 14 Ejemplo de dos niveles de descomposición de tareas

Es necesario prever la posible aparición de ataques por inyección de *prompts* y la generación de errores compuestos. Entre otros, se deberían establecer controles que garanticen la separación estricta entre datos e instrucciones, la correcta etiquetación y trazabilidad del origen del contenido, la limitación de privilegios de las herramientas utilizadas, la validación y sanitización de las entradas en cada etapa del proceso (en particular de la información en la memoria persistente) mediante *guardrails*.

También se podría realizar una evaluación automática de las decisiones finales adoptadas por el agente, e incluyendo las decisiones parciales en puntos críticos susceptibles que generar errores compuestos y la inclusión de mecanismos de evaluación de la confianza en la memoria persistente.

- ***Catálogo y listas blancas de servicios***

Se trata de disponer de un catálogo de servicios, que pueden incluir distintos LLMs, en el que se identifiquen versiones y, en particular, la fiabilidad que se da a cada servicio y/o adecuación para distintos contextos, así como evitar el efecto de alucinaciones que realicen llamadas a servicios inexistentes.

Esto permite que, para distintos contextos, se utilice como lista blanca, con flexibilidad para el uso de sistemas de IA agéntica en distintos tratamientos (para los desarrolladores, en caso de que las llamadas a funciones esté predefinida, o para limitar las herramientas invocables por el LLM en caso contrario). El catálogo debería abarcar servicios externos pero también internos, por ejemplo, repositorios de datos o servicios que permiten el acceso a la pantalla del operador.

- ***Limitación de servicios accesibles***

La limitación de servicios accesibles podría complementar al catálogo anterior por tratamientos específicos. De esta forma, cada tratamiento tendría políticas definidas sobre el máximo tipo de herramientas y accesos a datos que necesitaría para completar las tareas. Por ejemplo, en operaciones de consulta de normativa no se requeriría el acceso a servicios web si se dispone a priori de una recopilación actualizada disponible en forma de RAG.

- ***Control en la ejecución de herramientas***

La invocación de herramientas y servicios en Internet son *de facto* salidas parciales de la IA agéntica que pueden ser transparentes para las personas usuarias. Los controles que se podrían establecer sobre ellas son:

- Control de los parámetros con los que se invocan las herramientas, implementando *guardrails* y formatos rígidos para detectar parámetros erróneos o sesgados.
- Control de la respuesta de las herramientas, implementando nuevos *guardrails* sobre el contenido de entrada.
- Obligación de una acción humana de supervisión con relación a determinadas herramientas que pudieran tener un mayor impacto.

- ***Criterios y puntos de control para la intervención humana***

Desde el diseño se deberían definir criterios y puntos de control significativos o límites de actuación que requieran aprobación humana, especialmente antes de que se ejecuten acciones sensibles. Esto puede incluir:

- Acciones y decisiones de alto impacto, por ejemplo, la edición de datos sensibles, decisiones finales en ámbitos de alto riesgo (como la atención sanitaria o el ámbito jurídico), o acciones que puedan generar responsabilidad legal. Otro ejemplo podría ser la utilización de credenciales de usuario obtenidas de la memoria, solicitar una autorización previamente a su uso en subtareas críticas.
- Acciones irreversibles, por ejemplo, la eliminación permanente de datos, el envío de comunicaciones o la realización de pagos.
- Comportamientos atípicos o fuera de lo común, por ejemplo, cuando un agente accede a un sistema o base de datos fuera de su ámbito de trabajo, o cuando selecciona una ruta de entrega que duplica la distancia mediana.
- Definidas por el usuario. Los agentes pueden actuar en nombre de usuarios con diferentes niveles de tolerancia al riesgo. Además de los límites definidos por la organización, puede ofrecerse a los usuarios la opción de establecer sus propios límites, por ejemplo, exigir aprobación para compras que superen una determinada cantidad.

- ***Reversibilidad de las acciones de los agentes de IA***

Evaluar la necesidad de implementar medidas que permitan revertir determinadas acciones, por ejemplo si el agente puede modificar datos personales.

- ***Nivel de autonomía de acuerdo al tratamiento***

Incluir controles de la capacidad de autonomía ajustables para cada tratamiento que esté utilizando la IA agéntica, en función del impacto o por riesgo, desde ejecución autónoma en tratamientos de impacto y riesgo bajo hasta intervención humana obligatoria con gran granularidad en las operaciones del tratamiento cuando sean altos.

- ***Supervision humana efectiva***

Como es necesario, es necesario determinar cuándo integrar a expertos en puntos críticos del flujo para validar, refinar o anular las decisiones del agente (con los mecanismos de anulación oportunos) antes de que tengan un impacto real.

Para la evaluación de la intervención humana se aconseja tener en cuenta:

- Competencia y autoridad: tiene la autoridad o la tarea asignada que le permite alterar el resultado de la decisión automatizada.
- Preparación y formación: tiene la capacidad y las aptitudes para evaluar la decisión y los factores que determinan esa decisión con relación al contexto del tratamiento y el sistema automatizado empleado, en sus capacidades y limitaciones.
- Independencia: evaluar si existen presiones desde la organización o desde fuera de la organización que condicionen la disputa de la decisión por parte de la persona.
- Diligencia en el ejercicio de su competencia: en particular, si está sometido al sesgo de automatización.
- Medios para poder ejercer su competencia y cualificación.
    - Que existan los procedimientos y los medios técnicos para intervenir, en el momento o punto adecuado en tiempo y forma.
    - Que disponga de la información necesaria en tiempo y forma para poder ejercer su cualificación, en particular, conocer las consecuencias, riesgos de las decisiones en general, y las que se están tomando para los casos específicos y todos los aspectos que condicionan la decisión automatizada. Estos incluyen los datos del individuo concreto, pero también podrían incluir los procedimientos para la recogida de datos de entrada, los datos implícitos en el modelo que genera la decisión, los datos contextuales que no se han tenido en cuenta en la decisión automatizada, además de las capacidades y límites del sistema de decisión. También aquellos datos que la persona, en su cualificación, estime que son necesarios contemplar para el caso concreto y que no se han considerado en la decisión automatizada.
    - Que disponga de los recursos para poder ejercer su cualificación: las decisiones de la IA agéntica deben ser explicables por ejemplo, aplicaciones que le permitan analizar la información en el formato que se esté utilizando para la decisión automatizada, etc.
    - Que disponga del tiempo necesario para poder ejercer su cualificación para cada una de las decisiones que sean de su competencia.

- ***Rutas de escalamiento***

La supervisión humana podría complementarse con una monitorización automatizada en tiempo real para escalar cualquier comportamiento inesperado o anómalo. El escalamiento supone la implementación de protocolos y técnicas para

transferir el control de procesos automatizados a un operador humano cuando se detectan situaciones de alto riesgo, incertidumbre o anomalías.

Esto puede lograrse mediante la implementación de alertas para determinados eventos registrados (por ejemplo, intentos de acceso no autorizado a datos personales, o múltiples intentos fallidos de invocar una herramienta), el uso de técnicas de ciencia de datos para identificar trayectorias anómalas de los agentes, el uso de agentes para supervisar a otros agentes, el acceso a categorías especiales de datos cuando no es necesario, etc.

- ***Principio de los cuatro ojos***

En casos de procesos automáticos con gran impacto en los derechos y libertades de las personas, se puede plantear aplicar el principio de doble verificación por distintas personas, que constituyen una capa adicionales de confianza en el mecanismo de supervisión humana y fomentan la conciencia crítica del operador.

” y “. Trazabilidad”.

Cabe plantearse la necesidad de implementar un mecanismo ágil para gestionar un ciclo de vida en el consentimiento, donde el sujeto pueda decidir en cualquier momento modificar arbitrariamente sus demandas de tratamiento de datos, o revocar el consentimiento para el procesamiento o restringir el procesamiento en ciertos servicios.

Una medida podría ser determinar mecanismos para establecer la granularidad de dicho consentimiento, en cuanto a categorías de datos, categorías de tratamientos y categorías de destinatarios.

Podrían plantearse, en algunos tratamientos, el uso tanto de listas “blancas” como de listas “negras” que permitan definir con precisión las preferencias de los sujetos respecto a ciertas operaciones de tratamiento.

J. TRANSPARENCIA

El RGPD establece unas medidas mínimas de transparencia que son obligatorias.

Sin embargo, para superar un análisis de proporcionalidad o para reducir el riesgo es posible implementar medidas adicionales. Incluso, con el objeto de demostrar al sujeto que puede confiar las operaciones de tratamiento al sistema IA agéntico (como usuario, empleado, cliente, etc.) se podrían adoptar medidas como: información en tiempo real del flujo de datos, información sobre qué datos del sujeto se encuentran en los repositorios o servicios de terceros que están tratando los datos, accesos a registros de actividad de tratamientos y comunicaciones de datos, información sobre eventos intermedios en la cadena de razonamiento, contexto utilizado en el resultado, intervención humana realizada, posibilidad de pedir revisión o acción humana, acceso a certificaciones, auditorías o EIPDs de tratamientos.

K. ALFABETIZACIÓN

La alfabetización sobre sistemas de IA agéntica no solo es crucial para la eficiencia y eficacia de su implementación en tratamientos, sino que el conocimiento de las capacidades, fortalezas, debilidades y limitaciones de los mismos permiten una efectiva protección de datos personales. La alfabetización deberá tener en cuenta los diferentes roles que las personas tengan en el modelo de gobernanza o como usuarios en distintos tratamientos, y realizarse al menos a tres niveles:

- Nivel directivo, en el conocimiento necesario para que se tomen las decisiones adecuadas basadas en evidencias con relación a la inclusión de agentes de IA en tratamientos.
- Nivel de responsables TIC de desarrollo, contratación, despliegue, operación, mantenimiento y retirada de dichos sistemas, para que en particular se comprendan e identifiquen las implicaciones de protección de datos y las técnicas y medidas organizativas para implementarlas.
- Nivel de personas usuarias con distintos roles en los tratamientos en los que se utilicen sistemas de IA agéntica, con conocimientos de las posibilidades, implicaciones y limitaciones de estas herramientas.

En este proceso alfabetización un elemento clave es el DPD y los asesores de protección de datos en dos sentidos:

- Los DPD han de ser capaces de entender los fundamentos de las herramientas que se están utilizando, conocer las distintas alternativas técnicas y organizativas para implementar garantías y poder vislumbrar las oportunidades que para la protección de los derechos pueden ofrecer.
- Los DPD han de informar y asesorar al responsable o al encargado del tratamiento y a los empleados sobre la casuística de estos sistemas cuando se incluyen en un tratamiento y de supervisar que en su despliegue hay garantías de cumplimiento normativo.

VIII. REFLEXIONES FINALES

Los sistemas de IA, como puede ser la IA agéntica, han llegado para quedarse. Pretender ignorar su existencia, tanto desde el punto de vista de organización competitiva como si se trata de autoridades de control, supondría una pérdida de oportunidades estratégicas.

Conocer esta tecnología es necesario para tomar decisiones racionales sobre su implementación, decisiones basadas en evidencias. El conocimiento de una tecnología implica algo más que convertirse en usuario, sino comprender cuáles son sus fundamentos, sus implicaciones, sus limitaciones y la forma en que están implementadas. Tanto el rechazo irracional a todas las ventajas que presenta la IA agéntica, como el salto de fe para aceptar acríticamente cualquier tipo de implementación en tratamientos de datos personales podría ser perjudicial.

En particular, con un análisis objetivo, la implementación elegida de la IA agéntica permite más que asegurar la protección de datos, es decir, solo un enfoque reactivo antes las amenazas y vulnerabilidades. Una implementación de la IA agéntica teniendo en cuenta la protección de datos desde el diseño permite definir tratamientos basados en agentes que incorporen técnicas de mejora de la privacidad (Privacy Enhancing Technologies o PETs) que ofrezcan garantías superiores (o incluso que pudieran habilitar) a los tratamientos manuales. En sí misma, una IA agéntica puede ser un PET, si la utilizamos, por ejemplo, como herramienta para evaluar proactivamente los contratos y términos de servicio cambiantes de los proveedores que accede la organización.

En este aspecto la implicación del DPD y de los asesores de protección de datos son elementos clave. Para ello los DPD han de tener conocimiento de los tratamientos y de los principios de gestión de procesos, ser capaces de entender los fundamentos de las herramientas que se están utilizando, conocer las distintas alternativas técnicas y organizativas para implementar garantías y poder vislumbrar las oportunidades que para la protección de los derechos pueden ofrecer. Además, han de integrarse en las decisiones de diseño de los tratamientos y de los sistemas de agentes IA seleccionados para implementarlos, ya que las medidas posibles para gestionar el cumplimiento y el riesgo de la protección de datos están relacionadas con el cumplimiento de los demás objetivos y obligaciones de la entidad, y han de abordarse de forma integrada.

Para finalizar, nos encontramos con una tecnología que está en plena evolución y que requiere un análisis y experiencia, tanto de sus impactos, de sus medidas y de sus oportunidades para la protección de datos. Por lo tanto, consideren este texto como un estudio introductorio sin pretensiones de ser exhaustivo.

IX. REFERENCIAS

Reglamento (UE) 2016/679 (Reglamento General de Protección de Datos - RGPD) EUR-Lex - 02016R0679-20160504 - ES - EUR-Lex

Directrices del Grupo de trabajo sobre Protección de Datos del Artículo 29 sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679. (2017).

https://ec.europa.eu/newsroom/article29/items/612053/en

Agencia para la Ciberseguridad de la Unión Europea (ENISA). *Towards a framework for policy development in cybersecurity Security and privacy considerations in autonomous agents* (2018) https://www.enisa.europa.eu/publications/considerations-in-autonomous-agents

Lewis, P., et al. (2020). *Retrieval-Augmented Generation for Knowledge - Intensive NLP Tasks*. Publicado en NeurIPS: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Agencia Española de Protección de Datos (AEPD). Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial (2020) https://www.aepd.es/documento/adecuacion-rgpd-ia.pdf

Agencia Española de Protección de Datos (AEPD). Gestión del riesgo y evaluación de impacto en tratamientos de datos personales. (2021).

Comité Europeo de Protección de Datos. *Directrices 07/2020 sobre los conceptos de «responsable del tratamiento» y «encargado del tratamiento» en el RGPD* (2021) https://www.edpb.europa.eu/system/files/2023-10/edpb_guidelines_202007_controllerprocessor_final_es.pdf

Agencia Española de Protección de Datos (AEPD). Requisitos para Auditorías de Tratamientos que incluyan IA [ene 2021] https://www.aepd.es/documento/requisitos-auditorias-tratamientos-incluyan-ia.pdf

Yao, S., et al. (2022). *ReAct: Synergizing Reasoning and Acting in Language Models*. Publicado en ICLR 2023. https://arxiv.org/pdf/2210.03629

Agencia Española de Protección de Datos (AEPD). *Evaluación de la intervención humana en las decisiones automatizadas (2024)* https://www.aepd.es/prensa-y-comunicacion/blog/evaluacion-de-la-intervencion-humana-en-las-decisiones-automatizadas

Agencia Española de Protección de Datos (AEPD), "Introducción a LIINE4DU 1.0: Una nueva metodología para el modelado de amenazas para la privacidad y la protección de datos", (2024) https://www.aepd.es/guias/nota-tecnica-introduccion-a-liine4du-1-0.pdf

Future of Privacy Forum (FPF). (2024). *Minding Mindful Machines: AI Agents and Data Protection Considerations*. https://fpf.org/blog/minding-mindful-machines-ai-agents-and-data-protection-considerations/

Anthropic. (2024). *Model Context Protocol (MCP) Specification*. https://www.anthropic.com/news/model-context-protocol

Reglamento (UE) 2024/1689 de Inteligencia Artificial (RIA) https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A02024R1689-20240712

IBM. (s.f.). *What are AI agents? (2025)* https://www.ibm.com/think/topics/ai-agents.

Park, T. (2024). *Enhancing anomaly detection in financial markets with an LLM-based multi-agent framework*. 2403.19735

Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2025/2026). *AI Agents vs. Agentic AI: A Conceptual taxonomy, applications and challenges*. Information Fusion, Vol. 126, 103599. https://arxiv.org/pdf/2505.10468

OWASP Foundation. (2025). *Agentic AI-threats and mitigations*. https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/

Feng et al. *Levels of Autonomy for AI Agents* (2025) https://arxiv.org/abs/2506.12469

Infocomm Media Development Authority. *Model AI governance framework for agentic AI* Singapur (2026) https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf