Newsletter trimestral

CÁTEDRA

iDANAE

INTELIGENCIA · DATOS · ANÁLISIS · ESTRATEGIA

1T26

Marco de control de la IA agéntica

UNIVERSIDAD POLITÉCNICA DE MADRID

Management Solutions
Making things happen

Introducción

Se está produciendo un cambio de paradigma relevante en la inteligencia artificial (IA): la transición desde sistemas orientados principalmente a recuperar o sintetizar conocimiento (*search engines*) hacia sistemas capaces de planificar y ejecutar acciones (*action engines*) [1]. El ámbito empresarial está abordando una nueva fase de adopción de la IA, marcada por la construcción e integración de sistemas de agentes. A diferencia de un LLM tradicional, que es esencialmente pasivo y espera un *prompt*, un agente puede percibir, elaborar razonamientos, utilizar herramientas y ejecutar acciones complejas para cumplir un objetivo con una intervención humana mínima. Sin embargo, esta capacidad de acción introduce una paradoja operativa: cuanto mayor es la autonomía, mayor es el riesgo de resultados impredecibles.

Actualmente se observa una brecha en el mercado. Muchas empresas están desarrollando pruebas de concepto (PoCs) de agentes, pero todavía no se ha generalizado un despliegue masivo en entornos productivos de cara al cliente por la falta de garantías. Un LLM que alucina en un chat puede generar un riesgo operacional relativamente limitado[1], un riesgo legal o un riesgo reputacional; un agente que comete errores al ejecutar una acción (como una función de reembolso, un acceso a una base de datos, o un proceso en una planta de producción) constituye, en cambio, un riesgo corporativo crítico (podría producir daños materiales, generar un riesgo para las personas, o incurrir en acciones ilegales; el impacto económico de la acción ejecutada podría llegar a superar, potencialmente, los recursos económicos de la empresa).

Los sistemas de agentes presentan desafíos específicos que la ciberseguridad tradicional no siempre cubre por sí sola, como el no determinismo (al igual que en los LLM, una misma entrada puede generar salidas y acciones distintas), los bucles infinitos (con costes que pueden dispararse si el agente no logra resolver una tarea), o la manipulación del comportamiento del agente mediante instrucciones maliciosas en el *prompt* (*prompt injection*) o en el contexto (por ejemplo, el incidente del *chatbot* de un concesionario Chevrolet que llegó a aceptar la venta de un coche por $1 [2]).

Estos riesgos técnicos se traducen en retos operativos y organizativos de gran relevancia. La adopción de sistemas de agentes obliga a redefinir procesos, métricas de rendimiento y modelos de responsabilidad; a establecer vías seguras de escalado desde las PoCs hasta la producción; y a garantizar el cumplimiento normativo, el control de costes, la protección del dato y la trazabilidad de las decisiones automatizadas. Todo ello pone de manifiesto la necesidad de responder de manera estructurada e inteligente a estos nuevos retos, definiendo un marco de control robusto que permita a las empresas desplegar sistemas de agentes confiables, auditables y seguros. En esta publicación se repasan brevemente algunos de los estándares normativos relacionados con los sistemas agénticos, se presentan los principales componentes que forman parte de un marco de control de sistemas agénticos, y se plantean algunos retos que se han de abordar para el uso seguro de esta nueva tecnología.

[1] Desde un punto de vista teórico, el impacto económico derivado de un error producido por un LLM podría llegar a ser relevante. Por ejemplo, el uso de chatbots ha generado precedentes donde la justicia ha determinado que la empresa es legalmente responsable de lo que dice su IA en su web, lo que conlleva un riesgo legal claro, y cuyo impacto económico puede alcanzar, teóricamente, un nivel relevante para la empresa, dependiendo de cada caso. Hay sentencias donde se requiere una compensación económica a un cliente damnificado (véase, por ejemplo, [17]). A dicho impacto económico hay que sumar el coste que conlleva para la empresa el activar la revisión del propio sistema de IA, así como la preparación y la ejecución de su propia defensa legal, o el posible impacto reputacional. Por tanto, esta afirmación se ha de entender "en relación con el posible impacto económico que puede ocasionar un error producido por un agente o sistema de agentes", y no como una afirmación absoluta.

Regulación

La respuesta regulatoria al crecimiento de los sistemas de agénticos ha sido rápida. Las organizaciones deben alinear sus estrategias internas con estos estándares emergentes no solo para evitar sanciones, sino también para construir confianza. A continuación, se presentan tres marcos regulatorios que resultan especialmente relevantes en el panorama de la IA agéntica en 2026.

EU AI Act

La entrada en vigor del reglamento 2024/1689 por el que se establecen normas armonizadas en materia de IA en la UE (EU AI Act, [3]) plantea desafíos específicos para el despliegue de sistemas de agentes:

- IA de propósito general (*general-purpose AI*, o GPAI): el AI Act impone obligaciones específicas a los proveedores de modelos GPAI, como la documentación técnica, la política de copyright y la publicación de un resumen del contenido utilizado para el entrenamiento. Los despliegues de sistemas de agentes deben asegurarse de que los modelos subyacentes cumplen con estos requisitos cuando les resulten aplicables.
- Evaluación de conformidad para alto riesgo: no todo agente es, por sí mismo, un sistema de alto riesgo; la calificación depende del caso de uso y del ámbito en que se despliegue. No obstante, cuando un sistema de agentes entra en una categoría de alto riesgo, debe someterse a los requisitos y a la evaluación de conformidad correspondientes, incluyendo el requisito *human-in-the-loop* (HITL)

ISO/IEC 42001

La norma ISO/IEC 42001 constituye una referencia internacional para los sistemas de gestión de IA (AIMS). Frente a aproximaciones puramente locales o sectoriales, ISO/IEC 42001 proporciona un marco común para establecer, implementar, mantener y mejorar de forma continua la gobernanza de la IA [4].

Para los sistemas basados en agentes, la norma resulta especialmente útil por su énfasis en el ciclo de vida continuo y en la gestión sistemática de riesgos y oportunidades. En la práctica, este enfoque exige procesos formales para la gestión del cambio, la documentación, la trazabilidad y la mejora continua del sistema.

- Gestión del cambio: evaluar cómo nuevos datos, nuevas herramientas, cambios en *prompts* o ajustes del modelo pueden afectar al comportamiento del agente.
- Evaluación de riesgos e impactos: analizar, antes del despliegue y durante la operación, las consecuencias operativas, éticas y de cumplimiento asociadas al sistema.
- Trazabilidad y transparencia: documentar de forma suficiente el diseño, el funcionamiento y los controles del sistema, un aspecto crítico cuando los agentes toman decisiones o ejecutan acciones de manera autónoma.

Singapore Model AI Governance Framework for Agentic AI (MGF)

En enero de 2026, la Infocomm Media Development Authority (IMDA) de Singapur presentó, en el contexto del Foro Económico Mundial, un marco de referencia sobre la gobernanza de modelos de IA agéntica[2] [5]. Se trata de uno de los primeros marcos públicos de referencia específicamente orientados a los desafíos de gobernanza de los sistemas basados en agentes, más allá de los marcos centrados exclusivamente en IA generativa.

El marco propone una estructura de gobernanza basada en cuatro dimensiones críticas que las organizaciones deberían considerar:

1. Evaluación y acotación temprana del riesgo: determinar si la tarea es apta para la autonomía y establecer límites a las capacidades, accesos y ámbitos de actuación del agente.
2. Responsabilidad humana significativa: la autonomía técnica no elimina la responsabilidad legal y ética. El marco propone definir *checkpoints* obligatorios donde la ejecución se pausa para requerir aprobación humana explícita en decisiones de alto impacto (HITL).
3. Controles y procesos técnicos: incluye la implantación de medidas de evaluación, monitorización, identidad, permisos, pruebas y *red-teaming* específicas para someter a los agentes a escenarios adversos realistas.
4. Responsabilidad del usuario final: implica informar al usuario de que interactúa con un agente y dotarle de la información necesaria para comprender sus límites y utilizarlo de manera adecuada, reduciendo así el sesgo de automatización.

[2] En 2020 se había publicado la segunda versión del Model AI Governance Framework, orientada a la IA tradicional, y en 2024 se publicó el Model AI Governance Framework for Generative AI, un anexo para abordar los riesgos específicos de la creación de contenido (como las alucinaciones y la propiedad intelectual).

Principales componentes del marco de control para un sistema agéntico

El marco de control para un sistema agéntico ha de considerar múltiples componentes: (i) procesos de buena gobernanza; (ii) una correcta planificación y organización en el desarrollo del sistema; (iii) la incorporación de elementos técnicos y arquitecturales que incluyan barreras de defensa ante malfuncionamiento o un uso inadecuado; (iv) sistemas de seguimiento y monitorización que permitan identificar posibles errores desde el punto de vista del funcionamiento del sistema, el uso, el resultado esperado, así como sus impactos adicionales en términos del uso de recursos tecnológicos o de costes de operación; (v) o la posibilidad de incorporar la supervisión humana en el flujo de trabajo del sistema, entre otros. Estos componentes deben ser coherentes entre sí, y cubrir las distintas fuentes de riesgo que pueden detonar pérdidas económicas (derivadas de ámbitos operacionales, reputacionales, legales, etc.). Esta sección está dedicada a explorar algunos componentes de un marco de control ligados a elementos técnicos (por tanto, no se profundiza en elementos de gobernanza ni organizativos):

1. El diseño de la arquitectura tecnológica del sistema, que permite la incorporación de controles y cortafuegos para la evaluación y detección de errores.
2. El establecimiento de guardarraíles aplicables durante los procesos de ejecución del sistema.
3. La medición de métricas de rendimiento (tanto de los modelos como del sistema global) y establecimiento de mecanismos que permitan asegurar un comportamiento controlado o la parada del sistema ante un fallo.
4. La incorporación de mecanismos que permita controlar posibles impactos adicionales derivados de la presencia de errores (por ejemplo, en términos de consumo ineficiente de recursos tecnológicos o el incremento de los costes).
5. Por último, la incorporación de la supervisión humana durante la ejecución del proceso, tanto en la toma de decisiones críticas como en la aceptación de resultados (intermedios y finales).

A continuación, se exponen brevemente cada uno de ellos.

La arquitectura como elemento de control

El diseño arquitectónico es un elemento determinante para garantizar el control sobre sistemas de agentes. La ingeniería de sistemas de la IA agéntica converge hacia arquitecturas modulares, en las que el desacoplamiento de funciones como la planificación, la gestión de memoria (a corto y largo plazo) y el uso de herramientas permite una orquestación más robusta [7]. Esta transición no constituye solo una evolución técnica, que permite la intercambiabilidad de modelos, sino que también facilita el incorporar mecanismos y estructuras de control del sistema: al hacer más explícita la separación entre las funciones de orquestación, memoria, uso de herramientas y ejecución, estas arquitecturas favorecen la trazabilidad, la evaluación, el análisis de resultados a nivel de componente, y la aplicación de cortafuegos para evitar propagación de errores o establecer líneas de defensa ante ataques o mal funcionamiento. Asimismo, esto implica que, desde el punto de vista de la gobernanza, se puede realizar una asignación interna de responsabilidades más clara, en línea con los principios de gobernanza, control y *accountability* promovidos por el EU AI Act, la ISO/IEC 42001, o el Singapore Model AI Governance Framework for Agentic AI.

En la práctica, estas arquitecturas se materializan en una separación funcional en capas, cada una con responsabilidades específicas y puntos de control bien definidos. A continuación, se describe un esquema de cinco capas para un sistema de agentes[3]:

1. **Capa de interfaz y percepción**: es la frontera del sistema. Su responsabilidad va más allá de la simple comunicación: actúa como el primer cortafuegos cognitivo. Su función crítica es recibir los *inputs* del usuario (el objetivo que debe cumplir el sistema) y entregar *outputs*, filtrando activamente *prompt injections* antes de que alcancen el núcleo. Entre los componentes de control típicos de esta capa se encuentran los API *gateways*, los guardarraíles y la sanitización de datos.
2. **Capa de orquestación y planificación**: constituye el núcleo estratégico y de control del sistema. Su función primordial es abstraer la tarea: en lugar de ejecutar acciones de forma atómica, descompone los objetivos de alto nivel en un plan o flujo estructurado de ejecución. Esta capa coordina la activación selectiva de herramientas o subagentes especializados y preserva la coherencia operativa mediante mecanismos de gestión de estado. Entre sus componentes

[3] Por simplicidad, en este documento se muestra una visión reducida. Para una discusión completa de esta arquitectura, distintas opciones y casos de uso de despliegue, véase [6].

habituales se encuentran los planificadores, los enrutadores y los mecanismos de supervisión y recuperación ante errores [8].

3. **Capa de núcleo de agentes:** es la capa en la que residen los agentes. Cada agente es una unidad autónoma encapsulada con un propósito específico (rol), un sistema de instrucciones (*system prompt*) y un modelo de lenguaje (LLM) asignado. Su función es realizar una ejecución focalizada de las tareas que le son encomendadas.
4. **Capa de herramientas y servicios**: dota a los agentes de capacidad de impacto real e interacción con el mundo exterior. Sin estas herramientas, un agente es solo un chat; con ellas, es un sistema capaz de operar. En esta capa son imprescindibles componentes de control como el *sandboxing* (entornos de ejecución aislados) y la incorporación de mecanismos de revisión humana (*human-in-the-loop*) para acciones críticas.
5. **Capa de memoria y conocimiento**: esta capa gestiona la continuidad operativa del agente y el acceso a contexto relevante a lo largo del tiempo. Entre sus componentes habituales se encuentran la memoria a corto plazo —vinculada al historial y al estado de la sesión— y la memoria a largo plazo —implementada a menudo mediante mecanismos de recuperación externa, como *Retrieval-Augmented Generation* (RAG) apoyado en *embeddings* y búsqueda vectorial—. En esta capa resulta además crítico establecer controles de gobierno del dato para evitar la retención o recuperación indebida de información confidencial, mediante políticas de ciclo de vida del dato y mecanismos de filtrado en la fase de recuperación (*retrieval*), en línea con el principio de privacidad desde el diseño [9]. Una gobernanza robusta en esta capa garantiza que el historial de sesión y el conocimiento corporativo recuperado no se conviertan en vectores de fuga de información confidencial o de datos personales no autorizados, en línea con las obligaciones de seguridad y precisión del dato exigido en entornos regulados, como el marco de la Unión Europea para los proveedores de sistemas de IA de alto riesgo.

Este diseño por capas evita que un fallo en un punto colapse todo el sistema. Si un componente falla, las capas de orquestación o de herramientas pueden interceptar el error, contribuyendo a un sistema más resiliente y tolerante a fallos.

Adicionalmente a todo ello, en el diseño de la arquitectura se han de considerar estructuras de seguridad, autenticación y perfilado de usuarios. Por ejemplo, un enfoque habitual es el uso de paradigmas "Zero Trust", donde no se presume confianza implícita en usuarios, dispositivos o servicios por su ubicación en la red, ya sea interna o externa, sino que se exige autenticación, autorización y controles de acceso estrictos sobre los recursos del sistema [6].

Guardarraíles dinámicos

En el contexto de agentes autónomos, los mecanismos de control deben ir más allá de la configuración inicial del sistema, sino que deben operar de forma continua durante su ejecución, acompañando cada decisión y cada acción de este. La incorporación de agentes autónomos introduce un nuevo vector de riesgo crítico: la capacidad de acción. Investigaciones recientes alertan de que los sistemas de agentes son susceptibles al riesgo de *prompt injections* y propagación interagente de instrucciones maliciosas: un agente comprometido puede inyectar instrucciones maliciosas en la memoria de otro, creando una cascada de fallos sistemáticos dentro del sistema [10].

Para mitigar este riesgo no basta con instruir bien al modelo. Es necesario implementar guardarraíles dinámicos: capas de software determinista que envuelven al modelo probabilístico. Cuando el sistema de IA no tiene capacidad intrínseca para validar por sí mismo, en tiempo real, la corrección, la veracidad o la aceptabilidad de sus acciones, dicha validación se delega en un conjunto de controles externos que interceptan el tráfico antes y después de cada inferencia.

Desde un punto de vista operativo, y en consonancia con los marcos de gestión de riesgos y de seguridad como el NIST AI Risk Management Framework (AI RMF) [11], incluyendo su actualización para el perfil de IA generativa [12], o el OWASP Top 10 para LLM [13], estos controles pueden organizarse en cuatro ámbitos complementarios de defensa que permiten gestionar los riesgos de forma granular: datos, modelo, aplicación e infraestructura (véase tabla 1).

- **Guardarraíles de datos**: se orientan a asegurar la calidad, integridad, privacidad, procedencia, minimización de la información y protección de los datos para evitar el envenenamiento de los datos (*data poisoning*) que alimentan al sistema, incluyendo entradas del usuario, contexto recuperado, memoria externa y, en su caso, datos empleados en ajuste o evaluación. Mediante validación, saneamiento, clasificación y aplicación de controles de privacidad, estos mecanismos reducen la exposición de información sensible y refuerzan la fiabilidad del contexto sobre el que el agente genera sus respuestas.
- **Guardarraíles de modelo**: se centran en supervisar la seguridad y robustez del núcleo generativo. Incluyen mecanismos de moderación, validación de entradas y salidas, detección de *prompt injections*, *jailbreaks*[4], comprobaciones de coherencia y, en algunos casos, evaluadores basados en otros modelos (como "llm-as-a-judge") que pueden supervisar parámetros y/o resultados del sistema. Su objetivo es reducir respuestas inseguras, fuera de alcance o manifiestamente erróneas antes de que lleguen al usuario o desencadenen acciones posteriores.
- **Guardarraíles de aplicación**: operan sobre la lógica de negocio, restringiendo qué puede hacer el agente dentro de los límites de resiliencia y observabilidad, así como su interacción con herramientas y APIs. Incluyen la validación de parámetros, la aplicación de políticas sobre contenidos y acciones, la limitación de permisos, la validación estructural de salidas y, cuando procede, la incorporación de puntos de *human-in-the-loop* para operaciones sensibles. Asimismo, la selección dinámica de herramientas y servicios externos exige controlar qué servicios puede invocar un agente de forma autónoma, a fin de preservar el control sobre el tratamiento de datos, la trazabilidad y la asignación de responsabilidades.
- **Guardarraíles de infraestructura**: son mecanismos de seguridad y protección física y lógica en la plataforma donde se ejecuta el sistema, ya sea en la nube o en entornos on-premise. Incluyen controles de acceso, identidad y privilegios, segregación de entornos, aislamiento de procesos y herramientas mediante técnicas como *sandboxing*, así como monitorización y detección de anomalías. Su finalidad es reducir riesgos como el acceso no autorizado, la escalada de privilegios o la fuga de datos.

Métricas de rendimiento y mecanismos de actuación ante el error

Más allá del diseño de arquitectura, se requiere que, una vez desplegado en producción, un sistema agéntico disponga de mecanismos de monitorización y observación profunda (conocida como "observabilidad"). La incorporación de métricas granulares permite detectar ejecuciones ineficientes o la

[4] Ataques de inyección de instrucciones (*adversarial prompt injection*) diseñado para eludir los filtros de seguridad y las salvaguardas éticas integradas por los desarrolladores. En sistemas agénticos, el peligro de un jailbreak reside en que un agente del sistema ordene algo indebido, o que el sistema ejecute algo indebido (como, por ejemplo, lanzar código malicioso o borrar una base de datos), dado que está conectado a herramientas ajenas al agente o al sistema agéntico.

Tabla 1. Ambitos complementarios de Defensa

Ámbito de Defensa	Riesgos mitigados	Controles Clave (ISO 42001 / NIST / Singapur)	Impacto en Responsabilidad Legal
Datos	▸ Fuga de datos personales (PII: *Personally Identifiable Information*) ▸ Sesgos ▸ *Data poisoning*	▸ Minimización de datos ▸ Saneamiento de contexto (RAG) ▸ Auditoría de procedencia	Cumplimiento del RGPD y deber de diligencia en la gestión de activos informativos
Modelo	▸ *Jailbreaks* ▸ Alucinaciones ▸ Respuestas fuera de alcance	▸ Moderación de entradas/salidas ▸ Evaluadores externos (*LLM-as-a-judge*) ▸ Técnicas de *alignment*	Evidencia de supervisión técnica para mitigar la negligencia algorítmica
Aplicación	▸ Ejecución de acciones no autorizadas ▸ Errores contractuales	▸ Puntos de aprobación humana ▸ Validación de parámetros de API ▸ Guardarraíles a nivel de sistema	Protección frente a la formación de contratos electrónicos erróneos o vinculación involuntaria
Infraestructura	▸ Escalada de privilegios ▸ Acceso no autorizado a sistemas críticos	▸ *Sandboxing* de agentes ▸ Aislamiento de procesos ▸ Monitorización de anomalías de red	Prevención de daños a terceros y cumplimiento de estándares de ciberseguridad industrial

presencia de anomalías o errores, así como activar mecanismos de actuación controlada ante la identificación de un error. La integración de estos elementos está también orientada a hacer más robustos los atributos de mantenibilidad y analizabilidad definidos en el modelo de calidad para sistemas de IA de la norma ISO/IEC 25059:2023 [14].

A continuación, se muestran cuatro mecanismos que mejoran la detección de anomalías y la gestión del error:

1. La medición de la latencia del modelo y del sistema. Esto permite identificar posibles funcionamientos erróneos, asignación ineficiente de recursos, o sobrecargas. Se pueden incorporar las siguientes métricas:
    a. Rendimiento del modelo: mide la velocidad bruta del motor cognitivo. Se incluyen *time to first token* (TTFT, crítico para la percepción de latencia) y *time per output token* (TPOT, crítico para el rendimiento total).
    b. Rendimiento del sistema: captura la sobrecarga de la arquitectura del agente. Incluye la latencia de orquestación, los *spans* (segmentación temporal de cada operación que realiza el agente) y la latencia de las herramientas.
2. La monitorización del consumo de *tokens*, el establecimiento de límites de uso, la selección dinámica del modelo o la optimización de la arquitectura que permita reducir llamadas innecesarias, constituyen mecanismos adicionales para un mejor funcionamiento del sistema.
3. La incorporación de sistemas de *fallback* y gestión de errores: dado que los sistemas de agentes son probabilísticos y presentan dependencias externas, alucinaciones y flujos largos de ejecución, es muy importante que estén diseñados para fallar de forma controlada (en línea con el principio de resiliencia del NIST AI RMF [11]). De lo contrario, estos errores podrían bloquear el sistema, provocar bucles infinitos o generar respuestas incorrectas. Algunos mecanismos habituales son los reintentos controlados, los *fallbacks* funcionales (reglas deterministas en caso de error) o los *circuit breakers* (aislamiento de un componente del flujo de ejecución cuando falla).
4. El entendimiento de las decisiones: evaluar la secuencia de decisiones de los agentes es esencial, dado que su funcionamiento no determinista en la toma de decisiones o en la generación de resultados hace insuficientes los tests estándar por sí solos. Esto implica examinar la calidad de las decisiones, la solidez del proceso de razonamiento y la coherencia del resultado general. Analizar la trayectoria

incluye evaluar los pasos que el agente emplea para alcanzar su objetivo, como la selección de herramientas, las estrategias seguidas y la eficiencia con la que ejecuta cada tarea.

En última instancia, estos mecanismos no son solo un requerimiento técnico de control, sino una salvaguarda jurídica, puesto que ayudan en la reconstrucción del proceso de razonamiento del sistema ante un incidente, facilitando la explicabilidad, y permitiendo determinar si un error fue fruto de una entrada maliciosa del usuario, de un fallo estructural en la lógica de planificación del agente, o de modificaciones incorporadas previamente introducidas de forma maliciosa.

Mecanismos de gestión de impactos

A diferencia del software tradicional o de las consultas directas a un LLM, donde el impacto de una acción es relativamente predecible, los sistemas basados en agentes introducen una variabilidad mucho mayor. Ante un problema complejo, un agente autónomo puede decidir ejecutar múltiples llamadas a herramientas o iterar muchas veces antes de alcanzar una solución final, lo cual puede conllevar un uso inadecuado o ineficiente de recursos, así como a un incremento de los costes asociados. Sin mecanismos de control estricto, anomalías como bucles infinitos pueden agotar presupuestos en poco tiempo. Para mitigar este riesgo operativo, se pueden incorporar diversas soluciones técnicas:

- Trazabilidad y atribución de costes: la incorporación de plataformas de previsión, trazabilidad y control de costes que complementen los sistemas de facturación de los proveedores de modelos. Estas plataformas permiten etiquetar cada petición, desglosar el consumo de tokens y atribuir el gasto exacto a cada agente.
- Routing dinámico de selección de modelo: no todas las tareas requieren la capacidad de razonamiento de los modelos más avanzados y costosos. Las arquitecturas modernas implementan enrutadores dinámicos que evalúan la complejidad de la solicitud antes de realizar la llamada al modelo. Si la tarea es sencilla, el sistema deriva la petición a un modelo más pequeño, rápido y económico, reservando los modelos más potentes para razonamientos críticos.
- Optimización de arquitectura y uso de tokens: se deben codificar restricciones a nivel de orquestador, definiendo un uso máximo de tokens por tarea o sesión y diseñando una arquitectura robusta que reduzca pasos redundantes y evite razonamientos innecesarios.

Supervisión humana

A pesar del avance en los mecanismos automáticos de control, la intervención humana (human-in-the-loop o HITL) sigue siendo un elemento imprescindible en los sistemas de agentes. En entornos de alto riesgo o con acciones de alto impacto, el sistema debe incorporar mecanismos efectivos de supervisión y aprobación humana en puntos significativos del flujo de trabajo, por ejemplo, para la verificación de la corrección de un resultado (intermedio o final), o para la aprobación de parte de un proceso y la autorización para continuar hacia el siguiente paso. Esta supervisión no elimina la autonomía del agente, pero sí introduce un control proporcional al nivel de riesgo, autoridad y contexto operativo. Los marcos de gobernanza y regulaciones sobre la IA consideran la supervisión humana como un elemento clave en el desarrollo y operación de sistemas de IA responsables y seguros.

El patrón *human-in-the-loop* se puede entender como un protocolo de interacción que opera bajo dos premisas:

1. Validación de alto impacto: las acciones clasificadas como irreversibles o críticas, como transferencias bancarias o despliegue de código, requieren una validación humana.
2. Gestión de la incertidumbre: cuando la confianza del modelo cae por debajo de un umbral predefinido, el sistema escala el caso a un operador en lugar de generar una respuesta potencialmente errónea.

Desde un punto de vista técnico, estos aspectos se pueden resolver mediante marcos de orquestación (como LangGraph), que permiten establecer puntos de interrupción lógica[5]:

- Suspensión: el agente detiene su ejecución justo antes de la herramienta de acción, guardando todo su contexto en una base de datos.
- Intervención: un operador humano recibe una notificación, pudiendo aprobar, rechazar o modificar el estado.
- Reanudación: el agente retoma la tarea con el estado actualizado por el humano, ejecutando la acción como si él mismo hubiera llegado a esa conclusión.

La implementación de HITL convierte cada intervención humana en un activo de datos. Cada vez que un operador corrige a un agente, se genera un par de datos input-corrección de alta calidad. Estos datos pueden alimentar procesos posteriores de evaluación, ajuste o mejora del sistema, reduciendo con el tiempo la necesidad de supervisión intensiva.

[5] En algunos casos complejos, como la revisión de un modelo cuya confianza cae por debajo de cierto umbral, la aplicación de mecanismos de supervisión intervención puede no ser tan directa.

Retos empresariales

La convergencia entre capacidades técnicas avanzadas y nuevas exigencias regulatorias conlleva una serie de retos en el ámbito empresarial. La cuestión ya no es únicamente si los agentes funcionan, sino cómo integrarlos de forma segura, eficiente y alineada con el negocio.

En efecto, a medida que los agentes adquieren mayor autonomía, capacidad de planificación y acceso a herramientas externas, emergen nuevos desafíos en términos de control, cumplimiento, coste y gestión del riesgo. En este contexto, la evolución pasa por trasladar enfoques experimentales hacia marcos operativos sólidos que permitan escalar la autonomía de forma segura y medible.

A continuación, se exploran algunos de estos retos y formas de abordarlos.

1. Redefinición de procesos y nuevas métricas de rendimiento. Los procesos tradicionales no están diseñados para flujos no deterministas ni para agentes que planifican, invocan herramientas o ajustan su comportamiento en tiempo real. El reto consiste en rediseñar procesos clave para incorporar una autonomía controlada, así como en establecer nuevos KPI específicos (p. ej., ratio de escalado a HITL, coste por acción o estabilidad del planificador) que permitan evaluar valor, eficiencia y riesgo.

2. Escalar de PoCs a producción con garantías reales de seguridad y coste. Algunos proyectos pueden quedar atascados en pilotos que no escalan por miedo a alucinaciones con impacto operativo, costes imprevisibles o falta de auditoría. El reto consiste en industrializar un camino seguro hacia producción mediante arquitectura por capas, guardarraíles dinámicos y observabilidad, desplegando agentes con un coste controlado y una trazabilidad completa de decisiones y acciones.

3. Cumplimiento normativo y riesgo operacional en agentes autónomos

La llegada del EU AI Act, la ISO/IEC 42001 y los nuevos marcos internacionales para agentes exige que las empresas demuestren control sobre las decisiones automatizadas y sobre el uso de modelos GPAI. El reto consiste en diseñar un marco interno que asegure la conformidad, gestione evaluaciones de impacto, limite acciones críticas y documente la lógica de decisión sin frenar la velocidad de innovación.

4. Protección del dato y gestión segura de la memoria de los agentes. Los agentes que almacenan memoria persistente o emplean RAG pueden revelar información sensible a usuarios no autorizados si no existen controles estrictos. El reto consiste en garantizar un uso seguro de las memorias a corto y largo plazo, evitando fugas, atribuciones incorrectas o aprendizajes no deseados, y asegurando que solo se conserve la información necesaria para el rendimiento sin comprometer la privacidad.

5. Integrar la observabilidad y la trazabilidad como requisitos críticos del negocio. La autonomía de los agentes introduce incertidumbre operativa: decisiones no deterministas, secuencias largas de acciones y dependencias externas difíciles de predecir. Sin una monitorización adecuada, un error menor puede escalar hasta convertirse en un fallo sistémico o en un sobrecoste significativo. El reto consiste en dotarse de capacidades para monitorizar, registrar y explicar cada acción del agente en tiempo real, garantizando auditorías completas, detección temprana de anomalías y una operación segura y alineada con los objetivos del negocio.

6. Gestión dinámica del ciclo de vida. La capacidad de aprendizaje y adaptación de los sistemas basados en agentes plantea un desafío a los modelos de validación estáticos tradicionales. Es previsible que estos sistemas incorporen progresivamente mecanismos de aprendizaje continuo, adaptación y mejora iterativa basados en experiencia y evaluación, reduciendo en determinados casos la dependencia de intervención humana para tareas de ajuste y mantenimiento correctivo [15]. La auditoría deberá transformarse en un proceso continuo. Será necesario implementar mecanismos que aseguren que las adaptaciones del modelo no desvían su comportamiento de las políticas corporativas ni introducen sesgos no detectados en la fase de diseño.

7. Sistemas híbridos y certidumbre operativa. Para sectores altamente regulados, la naturaleza probabilística de los LLM sigue planteando retos de control y validación. En este contexto, están emergiendo arquitecturas híbridas que combinan la flexibilidad de razonamiento y orquestación de los LLM con componentes basados en reglas o modelos especializados para tareas concretas [16]. En determinados procesos, esta combinación permite imponer restricciones deterministas sobre acciones o resultados concretos, reforzando así robustez, la interpretabilidad y el control operativo en puntos críticos del flujo.

En conjunto, estos retos reflejan que la adopción de agentes no es únicamente un desafío tecnológico, sino un cambio en la forma de operar, gobernar y escalar sistemas de decisión automatizada.

Conclusión

El avance hacia sistemas agénticos marca un cambio estructural en la forma de diseñar, ejecutar y gobernar la automatización inteligente. La autonomía de estos sistemas abre un potencial transformador, pero también introduce riesgos operativos, éticos y regulatorios relevantes.

Para aprovechar su valor de forma segura, debe adoptarse un marco de control que combine arquitecturas modulares, guardarraíles dinámicos, capacidades avanzadas de monitorización y gestión de impactos de lo errores del sistema, así como la supervisión humana.

En este contexto, el gobierno del dato, la gestión del ciclo de vida y el cumplimiento normativo emergen como pilares esenciales para garantizar que los agentes operen con transparencia, resiliencia y alineamiento con los objetivos del negocio.

Un marco de control bien diseñado no limita la innovación, sino que la habilita, al proporcionar la confianza necesaria para que los agentes actúen como verdaderos aceleradores del rendimiento en múltiples tipos de tareas.

Bibliografía

[1] S. Nath, R. W. White, F. E. Faisal, M. E. Sharp, R. W. Gruen y L. R. Sivalingam, From Search Engines to Action Engines, Computer, vol. 58, nº 6, p. 59–68, June 2025.

[2] S. McGregor, Incident ID 622: Chevrolet Dealer Chatbot Agrees to Sell Tahoe for $1, AI Incident Database, 2023.

[3] Regulation (EU) 2024/1689 of the European Parliament and of the Council.

[4] International Organization for Standardization, Information technology - Artificial intelligence - Management system (ISO/IEC 42001:2023), 2023.

[5] Infocomm Media Development Authority (IMDA), Model AI Governance Framework for Agentic AI (Version 1)., Government of Singapore, 2026.

[6] National Institute of Standards and Technology, Zero Trust Architecture (NIST SP 800-207), U.S. Department of Commerce, 2020.

[7] J. Luo y al., Large Language Model Agent: A Survey on Methodology, Architectures, and Applications, arXiv:2503.21460, 2025.

[8] D. Souza y P. Machado, Toward Architecture-Aware Evaluation Metrics for LLM Agents, arXiv:2601.19583, 2026.

[9] Z. Zhang y al., A Survey on the Memory Mechanism of Large Language Model-based Agents, ACM Transactions on Information Systems, vol. 43, nº 6, pp. 1-47, April 2025.

[10] D. Lee y M. Tiwari, Prompt Infection: LLM-to-LLM Prompt Injection within Multi-Agent Systems, arXiv:2410.07283, October 2024.

[11] National Institute of Standards and Technology, AI Risk Management Framework (NIST AI RMF 1.0), U.S. Department of Commerce, 2023.

[12] National Institute of Standards and Technology, AI Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1), U.S. Department of Commerce, 2024.

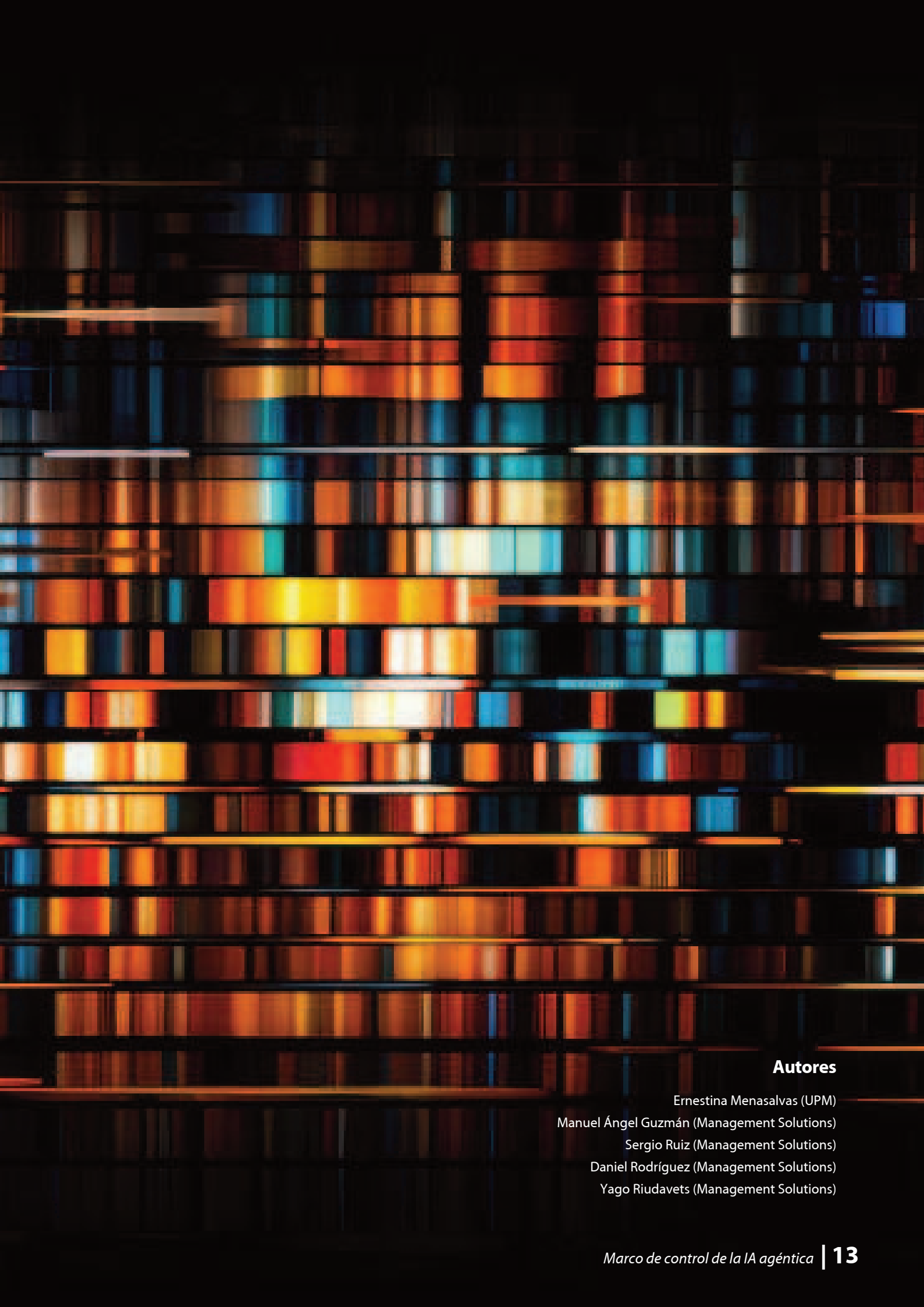
[13] OWASP, Top 10 for LLM Applications 2025, Version 2025.

[14] International Organization for Standardization, Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Quality model for AI systems (ISO/IEC 25059:2023).

[15] H. Gao y al., A Survey of Self-Evolving Agents: What, When, How, and Where to Evolve on the Path to Artificial Super Intelligence, arXiv:2507.21046, 2026.

[16] M. A. Farahania, M. I. Khana y T. Wuest, Hybrid Agentic AI and Multi-Agent Systems in Smart Manufacturing, 2026. [En línea]. Available: https://arxiv.org/pdf/2511.18258.

[17] Moffatt v. Air Canada. Case number BCCRT 149. Court: Civil Resolution Tribunal, British Columbia., 2024.



Autores

Ernestina Menasalvas (UPM)
Manuel Ángel Guzmán (Management Solutions)
Sergio Ruiz (Management Solutions)
Daniel Rodríguez (Management Solutions)
Yago Riudavets (Management Solutions)

UNIVERSIDAD POLITÉCNICA DE MADRID

La Universidad Politécnica de Madrid es una Entidad de Derecho Público de carácter multisectorial y pluridisciplinar, que desarrolla actividades de docencia, investigación y desarrollo científico y tecnológico.

www.upm.es

Management Solutions
Making things happen

Management Solutions es una firma internacional de consultoría, centrada en el asesoramiento de negocio, finanzas, riesgos, organización, tecnología y procesos, que opera en más de 50 países y con un equipo de más de 4.000 profesionales que trabajan para más de 2.200 clientes en el mundo.

www.managementsolutions.com

Para más información visita

blogs.upm.es/catedra-idanae/