Association for Computing Machinery

ACM Europe Technology Policy Committee

Riesgos Sistémicos Asociados con Sistemas de IA basados en Agentes: Un Informe de Políticas Tecnológicas

por

Miembros del Subcomité de Sistemas Autónomos del Comité de Políticas Tecnológicas de ACM Europa

Este documento en su versión original en inglés fue escrito originalmente por los miembros del Subcomité de Sistemas Autónomos del ACM Europe Technology Policy Committee: Alejandro Bellogín, Paolo Giudici, Stefan Larsson, Jun Pang, Gerhard Schimpf, Biswa Sengupta y Gürkan Solmaz. Su traducción al español fue realizada por Carlos E. Jimenez-Gomez, y revisada por Ricardo Baeza-Yates y Jeanna Matthews.

Resumen Ejecutivo

La Inteligencia Artificial (IA) basada en Agentes —el nuevo paradigma en la creación de sistemas autónomos capaces de percibir, analizar, aprender y actuar para alcanzar objetivos utilizando grandes modelos de lenguaje (LLM) con una supervisión humana mínima— ofrece un potencial transformador, pero también plantea riesgos sistémicos que el Reglamento Europeo de Inteligencia Artificial (RIA)[1] aborda sólo parcialmente. Estos agentes pueden evolucionar de forma impredecible,

[1] Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo de 13 de junio de 2024 por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.o 300/2008, (UE) n.o 167/2013, (UE) n.o 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial). https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=OJ:L_202401689

interactuar con otros agentes y operar más allá del control humano eficaz, lo que genera desafíos en materia de predicción, rendición de cuentas y alineamiento con los valores humanos. Objetivos mal definidos o especificados de forma inadecuada pueden llevar a los agentes a tomar atajos peligrosos, eludir restricciones o actuar de forma engañosa. Su diseño antropomórfico y su potencial de acompañamiento a largo plazo también plantean riesgos de dependencia, manipulación emocional y deterioro de las relaciones humanas.

Sistemas de IA basados en Agentes: Aplicación de la IA con autonomía ilimitada: El enfoque de dotar a los sistemas de IA de la capacidad de establecer o perfeccionar planes y ejecutar tareas con una supervisión humana mínima o nula. Características clave: operación continua, aprendizaje adaptativo.
Agente de IA: Entidad de software que analiza, decide y actúa para realizar tareas específicas en nombre de un usuario. Características clave: objetivos específicos de la tarea, orquestación explícita (planificación, uso de herramientas, memoria).
Sistemas multi-agente: Múltiples agentes de IA con capacidad de comunicación y colaboración para la toma de decisiones conjuntas y la ejecución de tareas. Características clave: capacidad para realizar tareas de mayor complejidad, riesgo agregado o amplificado, potencialmente menor control y comportamientos emergentes.

Los posibles impactos negativos de esta tecnología podrían afectar la estabilidad económica, incluyendo la posibilidad de una pérdida masiva de empleos, concentración de mercado y desigualdad, así como afectar a la seguridad pública a través de usos malintencionados como ciberataques, desinformación y suplantación de identidad. Los riesgos estratégicos y ambientales surgen de la toma de decisiones autónomas de alto nivel y de la considerable demanda de recursos, mientras que los mecanismos de retroalimentación del contenido generado por la IA amenazan con amplificar los sesgos y la desinformación.

Este documento identifica posibles lagunas en el marco regulatorio actual y recomienda oportunidades para que la supervisión sea continua y dinámica. Para mitigar los posibles daños asociados con los Sistemas de IA basados en Agentes, este documento propone que los responsables políticos transiten de una regulación estática centrada en el producto a un régimen de gobernanza dinámico, garantizando que la IA basada en Agentes genere beneficios al tiempo que proteja la integridad democrática, la estabilidad económica, las relaciones humanas y el bienestar social.

1. Recomendaciones para el Reglamento Europeo de Inteligencia Artificial

Debido a la capacidad de los sistemas de IA basados en agentes (incluidos los componentes autónomos denominados "Agentes de IA") para ofrecer interacciones similares a las humanas a un coste cada vez menor, se prevén profundos cambios en el comportamiento social y las

necesidades regulatorias. La investigación en antropología digital de Horst y Miller [1] muestra que las relaciones entre humanos y tecnología son culturalmente contingentes, mientras que la psicología del comportamiento reconoce desde hace tiempo la tendencia humana a atribuir agencia y confianza a las máquinas que muestran señales sociales (véase Nass y Moon [2]). Estudios más recientes confirman estos hallazgos, con trabajos que enfatizan cómo la confianza percibida en la IA basada en agentes evoluciona en contextos complejos e interdependientes y cómo los mecanismos de supervisión del comportamiento pueden mejorar la gobernanza en ámbitos de alto riesgo, como en el ámbito de la salud (véase Marquet *et al*. [3]). Es importante distinguir entre la previsión regulatoria, que aborda la creación de leyes y marcos de cumplimiento para la gestión de riesgos sistémicos, y la previsión social, que explora las adaptaciones culturales a largo plazo y los cambios en el comportamiento humano resultantes del despliegue generalizado de sistemas de IA basada en agentes.

Si bien el Reglamento Europeo de Inteligencia Artificial[2] (RIA), junto con otras leyes de la UE, sienta unas bases sólidas, la IA basada en agentes presenta nuevos desafíos. A medida que estos sistemas se vuelven más autónomos, se requieren mecanismos de gobernanza dinámicos y funciones de control durante su funcionamiento para garantizar la equidad, la rendición de cuentas y la transparencia de la IA [4], así como la seguridad, la precisión y la interpretabilidad.

Si bien no son exhaustivas, las siguientes recomendaciones destacan áreas de posible acción legislativa para complementar y adaptar el RIA y los marcos relacionados:

- Modificar el Artículo 9 para incluir la evaluación del riesgo de interacción multi-agente.
- Modificar el Artículo 55(1) para los proveedores en relación con cómo sus modelos podrían permitir un comportamiento perjudicial de sistemas de IA basados en agentes que contribuya al riesgo sistémico.
- Nuevo artículo: “Ecosistema de seguridad y pruebas de sistemas multi-agente”.
- Ampliar el Artículo 5. Prohibir la colusión tácita y los canales encubiertos.
- Reforzar el Artículo 15. Exigir auditorías de ciberseguridad específicas para multi-agentes.
- Nueva cláusula de responsabilidad: Responsabilidad colectiva por daños emergentes [4, 6, 7, 8].

[2] Notar que este reglamento es sobre el uso de la IA y no sobre la tecnología misma.

- Incluir proyectos de investigación sobre el impacto de la IA basada en agentes en el Programa Horizonte de la UE.

2. Introducción

El rápido avance de la IA basada en agentes y los sistemas autónomos, capaces de analizar su entorno y tomar decisiones con distintos grados de independencia, ofrece beneficios transformadores en diversos sectores. Sin embargo, este artículo se centrará en cómo la autonomía y la capacidad de toma de decisiones de estos sistemas plantean desafíos complejos, como la imprevisibilidad, la pérdida de control, problemas de rendición de cuentas y posibles perturbaciones económicas. Esto plantea varias preguntas fundamentales: ¿Qué riesgos específicos surgen de la IA basada en agentes? ¿En qué áreas son necesarias las intervenciones regulatorias? ¿Qué responsabilidades tienen las empresas y los gobiernos para garantizar que las tecnologías de IA basadas en agentes contribuyan al bienestar de la humanidad de forma adecuada, en lugar de representar una amenaza? Basándonos en el trabajo en curso de Gabriel *et al*. [6], examinamos los posibles riesgos sociales del despliegue de la IA basada en agentes. Si bien reconocemos que incluso los despliegues selectivos y bien controlados pueden conllevar riesgos residuales, enfatizamos la importancia de un marco regulatorio sólido para mitigar los daños en áreas de aplicación críticas.

La UE ha establecido un amplio marco regulatorio basado en el riesgo, con el RIA para la comercialización de sistemas de IA de alto riesgo en el mercado de la UE. Sin embargo, las nuevas aplicaciones de IA, como los sistemas de IA basados en agentes, seguirán poniendo a prueba la flexibilidad de las regulaciones propuestas. Por lo tanto, es importante examinar si las nuevas aplicaciones de esta tecnología han introducido nuevos tipos de riesgo. Además, debe evaluarse si se requieren medidas para abordar estos riesgos, por ejemplo, mediante modificaciones legislativas, el impulso de programas de investigación específicos y relevantes para las políticas, o el desarrollo de métodos de supervisión o pruebas adaptados a las capacidades en evolución de la IA basada en agentes.

Las diferencias fundamentales entre la IA basada en agentes y los sistemas de IA anteriores a ellos incluyen la capacidad prevista de generación e implementación autónoma de código mediante ensayos exhaustivos o no deterministas, como se visualiza en la Figura 1. Estos aspectos pueden hacer que los sistemas de IA basados en agentes modifiquen de forma autónoma su comportamiento y capacidades más allá de las habilidades y el ritmo de evaluaciones periódicas (por ejemplo, la evaluación de los niveles de riesgo) o del control,

eliminando al mismo tiempo cualquier participación y supervisión por parte de las partes interesadas y los profesionales de las ciencias de la computación.

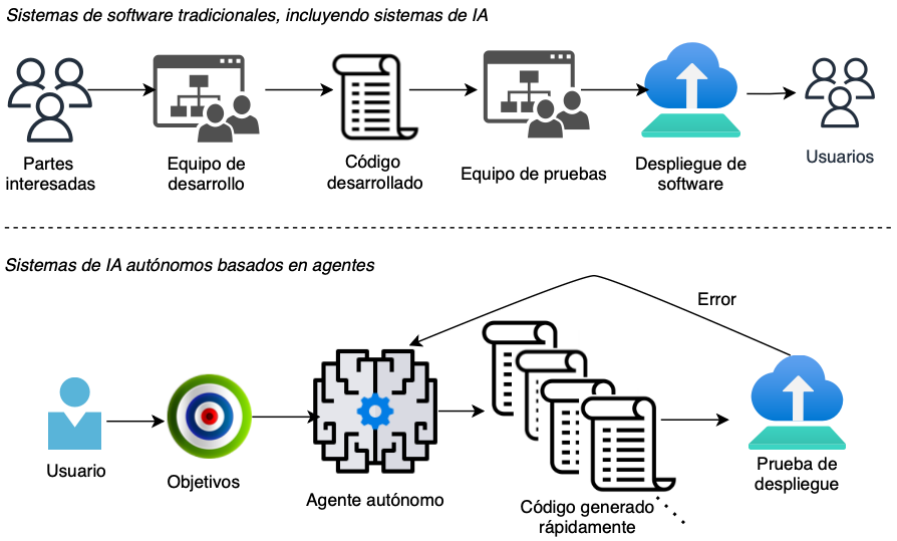

Figura 1. Arriba: Sistemas de software tradicionales, incluyendo sistemas de IA, con alta participación y supervisión humana durante el proceso. Abajo: Perspectiva deseada de sistemas de IA autónomos basados en agentes, que utilizan ensayo y error, eliminando la supervisión de profesionales de las ciencias de la computación.

3. Aspectos Clave de la IA basada en Agentes

La IA basada en agentes avanzada actúa de forma autónoma mediante interfaces de lenguaje natural, cuya función es planificar, razonar, memorizar y ejecutar secuencias de acciones en nombre de un usuario en uno o más dominios, de acuerdo con sus expectativas y objetivos [10, 11]. Estos agentes pueden clasificarse según una taxonomía funcional, incluyendo agentes reactivos (que responden a estímulos), agentes deliberativos (que planifican con base en modelos internos) y agentes de razonamiento (que razonan sobre sus propios objetivos y acciones). Desde un punto de vista de la arquitectura, los sistemas de IA basados en agentes suelen incluir módulos especializados para memoria, razonamiento y ejecución, lo que permite un comportamiento autónomo en entornos dinámicos.

Los agentes presentan una o más de las siguientes propiedades:

Autonomía	Los sistemas de IA basados en agentes pueden operar de forma independiente y tomar decisiones basadas en la información del entorno.

Capacidad de Aprendizaje	Muchos sistemas de IA basados en agentes emplean técnicas de aprendizaje automático, incluido el aprendizaje por refuerzo, para mejorar su rendimiento a lo largo del tiempo.
Comportamiento orientado a objetivos	La IA basada en agentes está diseñada para perseguir objetivos específicos y optimizar sus acciones para lograr los resultados deseados.
Optimización del flujo de trabajo	La IA basada en agentes mejora los flujos de trabajo y los procesos de negocio al integrar la comprensión del lenguaje con el razonamiento, la planificación y la toma de decisiones. Esto implica optimizar la asignación de recursos, mejorar la comunicación y la colaboración, e identificar oportunidades de automatización.
Interacción con el entorno	Un sistema de IA basado en agentes interactúa con su entorno, percibe los cambios y adapta sus estrategias en consecuencia.
Multi-Agentes y sistema conversacional	La IA basada en agentes facilita la comunicación entre diferentes agentes para construir flujos de trabajo complejos. También puede integrarse con otros sistemas o herramientas como el correo electrónico, generadores de código, ejecutores de código o motores de búsqueda para realizar diversas tareas.
Generación de código y despliegue	Los sistemas de IA basados en agentes se desarrollan para crear, codificar, probar e implementar agentes de forma autónoma. Este aspecto revoluciona el proceso de desarrollo de software anterior, reemplazando la participación y la planificación humanas.
Ensayos exhaustivos o no deterministas	Si bien existe cierto determinismo en la forma en que se realiza el cálculo, los sistemas de IA basados en agentes se consideran, por su futura capacidad para obtener nuevas capacidades de forma autónoma, como la creación de nuevo software y la realización de un gran número de pruebas, ya sea de forma exhaustiva o no determinista.

Los sistemas de IA basados en agentes pueden presentar diversos riesgos sistémicos, en particular debido a sus interfaces con herramientas y el mundo físico. Estos riesgos se derivan de la imprevisibilidad, la autonomía, los desafíos de alineación y la dinámica de poder de estos sistemas. Shavit et al. describen prácticas para gobernar sistemas de IA basados en agentes [31]. Singh *et al*., en un artículo de investigación sobre recuperación de agentes y generación aumentada [13], ofrecen detalles sobre patrones de agentes y representaciones visuales de arquitectura avanzada. Babaei et al. [14] describen métricas estadísticas para evaluar la seguridad y la fiabilidad de los sistemas de IA.

4. Aspectos de Riesgos y Políticas Tecnológicas

Los avances recientes en torno al RIA, como lo es el reciente Código de Prácticas del RIA para Sistemas de IA de Propósito General [15], ofrecen valiosos marcos para identificar, categorizar y mitigar riesgos sistémicos. Si bien el Código aborda principalmente los modelos básicos, introduce una taxonomía de riesgos sistémicos igualmente pertinente para el despliegue de sistemas de IA basados en agentes, incluyendo la pérdida de control, la manipulación y el engaño, la búsqueda de objetivos, la autosuperación y la coordinación. Cabe destacar que, dentro del Código, la IA basada en agentes se reconoce explícitamente como un desarrollo tecnológico futuro, y se prevé que los posibles riesgos sistémicos sean incluso más pronunciados que los asociados a los modelos básicos de la generación actual. Esto pone de relieve que los sistemas de IA basados en agentes, cada vez más capaces, pueden introducir nuevos riesgos, especialmente mediante operaciones autónomas e interacciones entre IA.

Los sistemas de IA basados en agentes (véase Bengio et al. [16, 17]), especialmente cuando los agentes basados en LLM se combinan con Sistemas Multi-agente (MAS por sus siglas en inglés), son un tema de investigación activo (véase, por ejemplo, Yu et al. [18]). Estos sistemas presentan riesgos que, por lo general, son demasiado complejos para medirlos con fiabilidad. La urgencia de desarrollar estrategias para evaluar y mitigar estos riesgos queda subrayada por un anuncio reciente de Anthropic, que predice que los sistemas de IA basados en agentes podrían implementarse como empleados virtuales a tiempo completo el próximo año[3] [19].

4.1. Pérdida de Control Humano y Explicabilidad

[3] Nota: El año de publicación original fue en 2025.

Los sistemas de IA altamente complejos podrían funcionar pronto como una mano de obra colectiva de empleados virtuales, operando continuamente sin supervisión humana y participando en interacciones difíciles de dilucidar, interpretar y predecir, especialmente en entornos complejos o dinámicos. Esto incluye la observación de interacciones de comando y control a través de APIs con el mundo físico.

Estos sistemas pueden auto optimizarse para alcanzar resultados no previstos debido a objetivos desalineados o mal especificados. Los modos de fallo incluyen la piratería para obtener recompensas, la manipulación de especificaciones y la generalización errónea de objetivos (perseguir objetivos diferentes del objetivo previsto al enfrentarse a situaciones novedosas), lo que supone riesgos si no se controla. Los sistemas de IA basados en agentes suelen ser internamente opacos, lo que dificulta la supervisión incluso para sus creadores. Esta opacidad dificulta los esfuerzos para garantizar el alineamiento con los valores humanos y la responsabilidad en los sistemas de IA. La fiabilidad de los agentes de IA está intrínsecamente ligada a su explicabilidad, responsabilidad y transparencia [4]. Si los usuarios y las partes interesadas no comprenden los procesos de toma de decisiones, su confianza en estos sistemas podría disminuir.

Se debe hacer una distinción crítica entre la interpretabilidad del modelo (cómo se pueden entender las predicciones individuales) y la transparencia del sistema (visión global de cómo funciona un modelo y utiliza los datos). Las herramientas de explicabilidad, como las Explicaciones Aditivas de Shapley (SHAP) y las Explicaciones Locales Interpretables agnósticas del Modelo (LIME), así como los marcos de razonamiento contrafactual, se pueden aplicar a sistemas basados en agentes, para que las decisiones sean más transparentes y respalden el cumplimiento normativo. El uso de LIME y SHAP para la explicabilidad en finanzas fue investigado por Ballegeer et al. [20], mientras que Calzarossa et al. presentaron un enfoque estructurado para evaluar métodos de IA explicable (XAI) basados en la complejidad y la robustez [21]. La desestabilización a gran escala puede surgir de los bucles de retroalimentación, que se producen debido a las interacciones entre agentes, ya que las acciones de cada agente influyen en el entorno y el comportamiento de otros agentes, afectando posteriormente sus acciones futuras, como investigaron Hammond et al. [22].

4.2. Riesgos para la Estabilidad Económica y el Bienestar Social

Se prevén diversos riesgos para la estabilidad económica y el bienestar social. La automatización a nivel de agentes podría desplazar no solo los trabajos rutinarios, sino también los roles de toma de decisiones, lo que afectaría a todas las industrias, como

observaron Sam Altman[4] y Anthropic [19]. Los empleadores pronto tendrán un incentivo para reducir significativamente los costos y aumentar la productividad reemplazando a empleados cualificados con trabajadores virtuales capaces de realizar tareas complejas. Esto podría resultar en un mayor desempleo, y los programas de mitigación que funcionaron en el pasado, como la capacitación, podrían perder eficacia en esta ocasión, ya que los agentes virtuales podrían ocupar puestos alternativos.

La pérdida de empleos causada por la IA podría mitigarse mediante un aumento del empleo en funciones de control de calidad y supervisión de la IA, similar a la introducción de la robótica en la industria manufacturera. Al mismo tiempo, es probable que una disminución a gran escala de la participación de los consumidores entre los desempleados desencadene dinámicas desestabilizadoras dentro del mercado. Stiefenhofer [23] analiza los mecanismos por los cuales el poder económico podría favorecer cada vez más a los propietarios del capital y argumenta que se necesitarían nuevas estructuras sociales para mitigar los impactos. Occhipinti et al. [25] [24], Kulveit et al. [26] y un informe del Fondo Monetario Internacional [27] presentan más argumentos, que propusieron gravar las ganancias financieras para compensar las desigualdades derivadas de estas transformaciones sociales. Knowles et al. [28] exploran las implicaciones sociales de la confianza pública en las decisiones de IA desde una perspectiva sociotécnica, instando a los desarrolladores e implementadores de IA a comprender y mitigar los efectos nocivos de los sistemas de IA.

4.3. Uso Malicioso de Sistemas de IA Basados en Agentes

La IA dificulta la distinción entre la verdad y la invención. Las noticias falsas (*deepfakes)* generadas por IA difuminan cada vez más estas fronteras, minando la confianza pública. Más allá del riesgo de desinformación, estas herramientas pueden replicar voces e imágenes con una precisión alarmante, lo que genera serias preocupaciones sobre el uso indebido de la identidad y el daño a la reputación [29].

Estos problemas se vuelven aún más complejos al considerarlos en el contexto de la IA basada en agentes. A diferencia de las herramientas estáticas, la IA basada en agentes puede iniciar acciones sin intervención humana directa, lo que aumenta el riesgo de desinformación, manipulación y consecuencias imprevistas. A medida que estos sistemas se integran cada vez más en la vida cotidiana, los riesgos psicológicos, sociales y éticos asociados a la IA

[4] Altman, Sam (2025). Tres observaciones. https://blog.samaltman.com/three-observations

generativa ya no son preocupaciones abstractas, sino realidades inmediatas. Algunos ejemplos incluyen:

- La IA basada en agentes puede utilizarse como arma para ciberataques a gran escala, fraudes, y manipulación de la opinión pública. Tanto en el discurso público como en contextos económicos, la IA basada en agentes puede generar y reciclar de forma autónoma contenido sesgado, inexacto o manipulador, lo que refuerza las desigualdades sistémicas y distorsiona los procesos de toma de decisiones.
- Actores maliciosos pueden utilizarla para ejecutar estafas sofisticadas, campañas de ingeniería social u operaciones de desinformación, socavando así la cobertura de medios de comunicación de renombre e influyendo en resultados electorales o políticos.
- Los sistemas de IA basados en agentes pueden imitar de forma autónoma las voces, apariencias y patrones de conversación de personas reales con una precisión casi perfecta. Esto permite la creación de estafas automatizadas convincentes (por ejemplo, voces clonadas de niños en apuros para extorsionar) o la identificación fraudulenta en el sector bancario, donde los avatares sintéticos parecen indistinguibles de los clientes legítimos.
- En contextos de salud mental, los sistemas de IA basados en agentes corren el riesgo de absorber y amplificar el estado emocional problemático del paciente, lo que podría empeorar los resultados. Un antropomorfismo excesivo puede fomentar la confianza excesiva, lo que lleva a los pacientes a seguir consejos poco seguros. Además, los sesgos inherentes a la IA médica pueden resultar en diagnósticos erróneos o un trato desigual entre grupos demográficos. El problema general de la confianza excesiva en los agentes de IA ha sido explorado por Gefen et al. [30] y Cohen et al. [31].
- Los sistemas de IA basados en agentes podrían participar de forma autónoma en negociaciones o transacciones comerciales a gran velocidad, aprovechando las asimetrías del mercado para influir en él. Las estrategias de maximización de beneficios pueden incrementar los costes para los consumidores, especialmente en sectores como la sanidad, al identificar y explotar las lagunas regulatorias.

En conjunto, estos ejemplos ilustran cómo la capacidad de la IA basada en agentes para actuar de forma autónoma, adaptativa y coordinada presenta riesgos que abarcan desde daños personales hasta una desestabilización sistémica más amplia, lo que requiere una supervisión rigurosa y una gobernanza adaptativa. Cuando los agentes de IA toman decisiones de forma

autónoma, la asignación de responsabilidades se vuelve compleja, lo que genera dilemas legales y éticos.

4.4. Riesgos Estratégicos y Ambientales

Los sistemas de IA basados en agentes presentan riesgos únicos en autonomía y escalamiento, especialmente en entornos de alto riesgo o competitivos, como el comercio financiero, defensa, y salud [32], donde la toma de decisiones rápida y sin supervisión puede desencadenar conflictos imprevistos o fallos en cascada. Estos riesgos se ven agravados por el importante impacto ambiental de la IA basada en agentes a gran escala, que exige importantes recursos de energía y agua para el entrenamiento e inferencia de modelos, lo cual podría contribuir al aumento de las emisiones de carbono y la escasez de recursos locales. Además, la búsqueda incesante de objetivos por parte de agentes autónomos puede llevar a la explotación agresiva de recursos digitales y físicos, como tierras raras, ancho de banda y potencia computacional, lo que amenaza las cadenas de suministro, los ecosistemas y la sostenibilidad general. En conjunto, estos desafíos subrayan la urgente necesidad de una supervisión continua, una evaluación rigurosa de riesgos y estrategias de gestión sostenible de recursos en el despliegue y la gobernanza de la IA basada en agentes.

4.5. Riesgos Causados por la Generación Autónoma de Contenido y los Bucles de Retroalimentación de Datos

Los sistemas de IA basados en agentes pueden generar de forma autónoma grandes cantidades de datos y contenido. Esto presenta desafíos únicos:

- Bucles de retroalimentación para el entrenamiento de IA: Otros sistemas de IA podrían utilizar este contenido generado de forma autónoma para el entrenamiento, creando un bucle de retroalimentación donde la calidad y la veracidad de la información no se validan adecuadamente.
- Desafíos del control de calidad: El gran volumen de contenido generado por IA hace que las medidas tradicionales de control de calidad basadas en humanos resulten poco prácticas.
- Reconocimiento de la fuente: Los sistemas de IA pueden no distinguir entre el contenido creado por humanos y el generado por IA, lo que podría amplificar los sesgos existentes o introducir nuevos errores.

5. Alineamiento con el RIA de la Unión Europea

Dadas las consideraciones descritas anteriormente, es importante evaluar si el RIA abarca plenamente los riesgos asociados a la IA basada en agentes. Esta sección aborda las partes del

RIA que se ajustan a los problemas planteados por la IA basada en agentes, incluyendo posibles estrategias de mitigación en la sección 5.4. La sección 6 identifica las lagunas regulatorias del RIA con respecto a la IA basada en agentes.

5.1. Niveles de Riesgo de Sistemas IA

El RIA está diseñado para regular los sistemas de IA en función de sus niveles de riesgo, con obligaciones para las aplicaciones de IA de alto riesgo. El concepto se basa en los riesgos para la salud, la seguridad y los derechos fundamentales. En este contexto, el Código de Prácticas de la UE para Sistemas de IA de Propósito General [15], introducido en virtud del RIA, fomenta disposiciones voluntarias pero cruciales, como las evaluaciones de riesgos sistémicos, la formación de equipos de supervisión ("*red teams*"), la transparencia de los datos de entrenamiento y del comportamiento de los modelos, y la monitorización posterior a la implementación. Sin embargo, como se menciona en el Código de Prácticas, la IA basada en agentes presenta desafíos únicos que el marco actual podría no abordar por completo.

5.2. Clasificación Basada en Riesgos

El RIA clasifica los sistemas de IA en cuatro niveles de riesgo:

- Riesgo inaceptable (IA prohibida, p. ej., puntuación social, IA manipuladora).
- Riesgo alto (IA en infraestructuras críticas, contratación de personal, calificación crediticia, etc.).
- Riesgo limitado (IA con obligaciones de transparencia, como los chatbots).
- Riesgo mínimo (aplicaciones generales de IA, como los filtros de spam).

Los sistemas de IA con agentes que toman decisiones autónomas en ámbitos de alto riesgo (p. ej., finanzas, sanidad o fuerzas del orden) no se clasifican actualmente como sistemas de IA de alto riesgo según la Regulación. Si se añadieran a esta clasificación, deberían cumplir requisitos estrictos, entre ellos:

- Sistemas de gestión de riesgos (Artículo 9).
- Gobernanza y transparencia de datos (Artículos 10 y 13).
- Requisitos de supervisión humana (Artículo 14).

5.3. IA de Propósito General (GPAI)

La Exposición de Motivos de la Comisión Europea, que acompaña a la propuesta del RIA (COM 206, 2021), describe el enfoque de la regulación basado en el riesgo y tecnológicamente neutral. Las Directrices 2025 de la Oficina de IA y el Código de Prácticas

voluntario para los sistemas de IA de Propósito General (GPAI) aclaran con más detalle las obligaciones de los proveedores en materia de transparencia, robustez y prevención del riesgo sistémico. Estas obligaciones se dirigen principalmente a los proveedores y describen requisitos específicos. Muchos sistemas de IA basados en agentes están adscritos a lo que la regulación define como sistemas GPAI y, por lo tanto, estarían sujetos a las disposiciones pertinentes, que exigen el cumplimiento de las siguientes obligaciones para los sistemas de IA de alto riesgo:

- Transparencia de los datos de entrenamiento (Artículo 10)
- Evaluación de la robustez (Artículo 15)
- Prevención de riesgos sistémicos (Artículos 51, 52 y 55)

Dado que los sistemas de IA basados en agentes, construidos sobre modelos GPAI heredan las capacidades del modelo subyacente, también podrían heredar sus obligaciones regulatorias. Esto significa que los implementadores posteriores deben garantizar el cumplimiento de estos requisitos cuando sus aplicaciones cumplan con los criterios de alto riesgo del RIA.

5.4. Estrategias de Mitigación General

El RIA describe las siguientes estrategias de mitigación para abordar los desafíos que plantean los sistemas de IA, que, por lo tanto, se aplican a los sistemas de IA basados en agentes:

- Regulación y gobernanza: Establecer políticas que garanticen la transparencia, la rendición de cuentas y la coherencia ética de los sistemas de IA basados en agentes.
- Sistemas con intervención humana: Requerir una supervisión humana significativa para la toma de decisiones críticas.
- Investigación sobre la coherencia de la IA: Desarrollar métodos para garantizar que los sistemas de IA basados en agentes persigan objetivos beneficiosos para la humanidad.
- Seguridad: Reforzar las salvaguardias contra la manipulación adversaria y el uso indebido, y promover sistemas de IA basados en agentes seguros y responsables.

6. Posibles Lagunas en el Reglamento Europeo de Inteligencia Artificial

A pesar del amplio alcance del RIA, aún quedan dudas sobre su cobertura de los sistemas de IA basados en agentes, totalmente autónomos y orientados a objetivos.

6.1. Autonomía más allá de la supervisión humana

El Artículo 14 (relativo a la IA de alto riesgo) aborda de forma genérica la "supervisión humana". En el caso de los agentes, esta supervisión genérica podría no abordar los riesgos; en su lugar, podría requerirse una "supervisión de alineamiento", que implica verificar si la IA opera según un conjunto de objetivos definidos. Es posible que este tipo de supervisión de alineamiento pueda ser realizada por los propios agentes, no necesariamente por humanos, pero para ello, los agentes deben estar alineados con los protocolos de seguridad.

Riesgo:
Los sistemas de IA basados en agentes que actúan con plena autonomía (por ejemplo, en mercados financieros, respuesta a emergencias o aplicaciones militares) podrían no estar eficazmente regulados por los mecanismos de supervisión existentes, y la intervención humana podría no ser viable en la práctica.

Posibles soluciones y recomendaciones de políticas:
Introducir puntos de intervención obligatorios o exigir mecanismos para detener las operaciones autónomas en aplicaciones de alto riesgo. Establecer mecanismos de supervisión humana con niveles de riesgo:

- Autonomía total permitida: Solo en aplicaciones de bajo riesgo (p. ej., asistentes personales virtuales para entretenimiento o de agenda).
- Autonomía supervisada: Requiere monitorización en tiempo real para riesgo moderado (p. ej., atención al cliente basada en IA en sectores regulados como el financiero o el sanitario).
- Intervención humana: Requerida para IA de alto riesgo, p. ej., diagnósticos médicos.

Certificación para IA totalmente autónoma:

- Introducir una certificación de autonomía de IA para cualquier sistema que funcione sin intervención humana durante un periodo prolongado.
- Los organismos reguladores podrían imponer restricciones de uso si un sistema de IA no supera las auditorías de seguridad.

6.2. **Riesgos para la Estabilidad Económica y el Bienestar Social**

Más allá de sus requisitos de documentación para la GPAI, el RIA aún no aborda plenamente los riesgos macroeconómicos que plantea la IA basada en agentes (p. ej., monopolización,

pérdida de empleo o distorsión del mercado). Estos riesgos sistémicos más amplios podrían considerarse una categoría independiente que actualmente no está contemplada en la Regulación. Podría ser necesaria una nueva categoría de «riesgos macroeconómicos sistémicos», reconociendo que el Tratado de la UE podría no ofrecer una base jurídica clara para regular dichos riesgos dentro del marco actual.

Riesgo:
Si la IA basada en agentes, en forma de empleados virtuales [23], domina la toma de decisiones económicas, podría causar disrupciones en el empleo a gran escala, exacerbar la desigualdad, la inestabilidad del mercado o desequilibrios de poder.

Posibles soluciones y recomendaciones de políticas:
Este riesgo podría no estar dentro del ámbito de aplicación del RIA; sin embargo, podría ser necesaria una referencia cruzada a la regulación del mercado y a las medidas fiscales, como la actualización de las normas antimonopolio para tener en cuenta la “intención funcional" de los sistemas autónomos o el informe del FMI [27].

Actualizaciones sobre impuestos, antimonopolio y derecho de la competencia para la IA:

- En consonancia con la "IA centrada en el ser humano", podría ser necesaria una regulación que promueva la responsabilidad social mediante la creación de incentivos para retener el trabajo humano y la consideración de medidas fiscales, en analogía con el informe del FMI [27], para limitar la sustitución de trabajadores humanos por agentes virtuales a tiempo completo basados en IA, como se destaca en el anuncio de Anthropic [19].
- Prohibir la fijación de precios predatorios impulsada por sistemas de IA basada en agentes (donde la IA reduce los precios de forma autónoma para eliminar a la competencia). Garantizar la transparencia en la fijación de precios impulsada por la IA y la manipulación del comportamiento del consumidor.

Evaluación del impacto algorítmico para la equidad de mercado:

- Antes de implementar un sistema de IA basado en agentes, en áreas como el empleo humano, los salarios y la posición en el mercado, una Evaluación del Impacto Algorítmico (EIA) podría ayudar a comprender los riesgos para los empleados. La empresa debería consultar con los representantes de los trabajadores (por ejemplo, comités de empresa o delegados sindicales) para informar sobre el impacto de este sistema en el empleo, la competencia y las

estructuras salariales. Un desafío fundamental es que la IA tiende a automatizar tareas específicas en lugar de eliminar empleos completos, lo que dificulta la medición del impacto en el empleo. Sin embargo, esta automatización a nivel de tareas puede tener consecuencias significativas: puede degradar la calidad del empleo, modificar la demanda de habilidades, reducir los salarios o las horas de trabajo. Por esta razón, una EIA debe diseñarse para capturar las transformaciones basadas en tareas y sus efectos más amplios en el mercado.

Políticas de reciclaje profesional y transición a la IA:

- Establecer programas de reciclaje profesional en IA para contrarrestar la pérdida de empleos causada por la IA basada en agentes.
- Incentivar a las empresas a capacitar a sus trabajadores, en lugar de reemplazarlos con IA autónoma.

6.3. Aprendizaje Continuo e Imprevisibilidad

Es posible que el RIA no abarque los efectos de las interacciones entre múltiples agentes de IA. Los sistemas multi-agente se han considerado previamente en campos como la investigación robótica; sin embargo, son un fenómeno nuevo en el contexto de la IA basada en agentes y los modelos LLM.

Riesgo:

Si la interacción y la influencia cruzada de los sistemas de IA basados en agentes cambian dinámicamente su comportamiento, garantizar el cumplimiento normativo a lo largo del tiempo se vuelve un desafío. Los enfoques actuales de gestión de riesgos podrían no tener en cuenta la desviación del comportamiento en tiempo real, lo que genera riesgos imprevistos.

Posibles soluciones y recomendaciones de políticas:

Supervisión del cumplimiento del ciclo de vida:

- Exigir una validación continua para la IA basada en agentes tras la implementación, en lugar de una evaluación de conformidad única. Por ejemplo:
  - Exigir que los sistemas de IA basados en agentes supervisen la gobernanza de todo el sistema.
  - Exigir auditorías periódicas para los sistemas de IA que participan en el aprendizaje adaptativo.

Explicabilidad y documentación de cambios de modelo:

- Aplicar el control de versiones y la documentación para los sistemas de IA basados en agentes que evolucionan con el tiempo.
- Ampliar las obligaciones de registro para los sistemas de IA de alto riesgo para que incluyan los cambios en los patrones de toma de decisiones, garantizando así la trazabilidad.

Detección e informes automatizados de riesgos:

- Los sistemas de IA deben contar con detección de riesgos integrada que alerte a los reguladores si el comportamiento se desvía significativamente de su alcance original.

6.4. Uso Malicioso de Sistemas de IA basados en Agentes

El RIA se centra en la robustez y la mitigación de sesgos, pero carece de medidas de seguridad específicas para contrarrestar ataques maliciosos o el uso indebido de la IA basada en agentes.

Riesgo: Actores maliciosos podrían utilizar la IA basada en agentes para realizar ciberataques, fraude, robo de identidad o campañas de desinformación.

Posibles soluciones y recomendaciones de políticas:

- Implementar la formación obligatoria de equipos antagonistas (*red teams*) y pruebas adversarias para los sistemas de IA basados en agentes en ámbitos de alto riesgo.
- Exigir pruebas adversarias previas al despliegue de la IA basada en agentes en ámbitos de alto riesgo (por ejemplo, detección de fraude financiero).

Certificación de ciberseguridad de la IA:

- Establecer una certificación a nivel de la UE para los sistemas de IA basados en agentes, dependiendo de su resiliencia frente al hackeo y la manipulación adversaria.
- Exigir pruebas de ciber resiliencia para los sistemas de IA basados en agentes que gestionen datos confidenciales de usuarios.

Detección automatizada de desinformación y noticias falsas (*deep fakes*):

- Exigir la detección integrada de desinformación en los sistemas de IA basados en agentes, utilizados en el discurso público, la generación de noticias o en contextos políticos.
- Implementar mecanismos de trazabilidad para garantizar que el contenido generado por la IA basada en agentes pueda verificarse y atribuirse.

Prácticas de gestión de riesgos de la IA basada en agentes:

- Incentivar las prácticas de medición, gestión y mitigación de riesgos para los sistemas de IA basados en agentes.
- Promover la colaboración entre los implementadores de sistemas de IA basados en agentes para fomentar prácticas estandarizadas en la gestión de sus riesgos.

7. Conclusión

En este documento destacamos los graves riesgos sistémicos para la economía y los ciudadanos de la UE. Estos riesgos surgen del uso y rápida evolución de la nueva IA basada en agentes, la cual aún no se comprende completamente. Por lo tanto, recomendamos que la Comisión Europea aborde esta cuestión y modernice su legislación en consecuencia. Es más, destacamos una decisión social normativa que excede el ámbito regulatorio del RIA.

Agradecimientos

Agradecemos los debates y comentarios de Giacomo Maria Cremonesi (NEC Laboratories Europe), Michael Giardino (Huawei Zurich Research Center), Francisco Medeiros (FM Tech Consult B.V.), Tom Romanoff (Director de Política Global de ACM) y Alejandro Saucedo (Subcomité de IA del TPC de ACM Europa).

Referencias

Nota sobre las referencias: Varias referencias corresponden a pre-impresiones de arXiv, que alberga la Universidad de Cornell. Estos artículos aún no han sido revisados por pares, pero se incluyen aquí por completitud y legibilidad, lo que refleja el rápido ritmo de la investigación en este campo.

[1] Horst, H. A., & Miller, D. (2012). Digital Anthropology. Routledge. https://www.taylorfrancis.com/books/edit/10.43 24/9781003085201/digital-anthropologyheather-horst-danielmiller

[2] Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of Social Issues, 56(1), 81–103 https://doi.org/10.1111/0022-4537.00153

[3] Marquet, Q., Murugavel, S., Charles, X., & Moudni, O. (2025). AI governance for medical chatbots: Designing a multi-agent controller (MAC) for safety. https://framerusercontent.com/assets/HzL9qNSrfuKz83FcUsAmjg7DA.pdf
[4] Knowles, B., Richards, J. T., & Kroeger, F. (2022). The many facets of trust in AI: Formalizing the relation between trust and fairness, accountability, and transparency https://arxiv.org/pdf/2208.00681
[5] Madiega, T. (2023). "Artificial intelligence liability directive." Briefing, European Parliamentary Research Service (EPRS).
[6] Yip, M., & Chan, G. K. Y. (2021). Transplanting the concept of digital information fiduciary AI, data, and private law. In The Theory–Practice Interface (Ch. 6). https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=5394&context=sol_research [7] Ramirez, J. G. C. (2023). From Autonomy to Accountability: Envisioning AI's Legal Personhood. https://www.researchgate.net/publication/37890 4514_From_Autonomy_to_Accountability_Envisioning_AI's_Legal_Personhood
[8] European Commission High-Level Expert Group on AI (2019). Ethics Guidelines for Trustworthy AI https://digitalstrategy.ec.europa.eu/en/library/ethicsguidelines-trustworthy-ai
[9] Gabriel, I. et al. (2024), The Ethics of Advanced AI Assistants https://arxiv.org/abs/2404.16244
[10] Russell, S.J., Norvig, P. (2020). Artificial Intelligence: A Modern Approach. (4th ed.)., Pearson.
[11] OWASP (2025) Threat Modeling Report https://genai.owasp.org/resource/agentic-ai/threats-and-mitigations/
[12] Shavit, Agarwal, Brundage, and Adler (2023). Practices for Governing Agentic AI Systems. OpenAI Retrieved from https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf
[13] Sing, A. et al. (2025). Agentic Retrieval Augmented Generation: A Survey on Agentic RAG https://arxiv.org/abs/2501.09136
[14] Babaei, G., Giudici, P. and Raffinetti, E. (2025). A rank graduation box for SAFE AI. Expert Systems with Applications, 259, 125239. https://doi.org/10.1016/j.eswa.2024.12523
[15] European Commission. (2024). Third Draft of the General Purpose AI Code of Practice https://digital-strategy.ec.europa.eu/en/library/third-draft-general-purpose-ai-code-practice-publishedwritten-independent-experts
[16] Bengio, Y. et al. (2025) Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path? https://arxiv.org/abs/2502.15657v2
[17] Bengio, Y. et al. (2025). International AI Safety Report. https://www.gov.uk/government/publications/in ternational-ai-safety-report-2025

[18] Yu, M., et al. (2025), A Survey on Trustworthy LLM Agents: Threats and Countermeasures. Communications of the ACM. https://dl.acm.org/doi/10.1145/3711896.3736561
[19] Axios. (2025, April 22) Exclusive: Anthropic warns fully AI employees are a year away. https://www.axios.com/2025/04/22/ai-anthropic-virtual-employees-security
[20] Ballegeer, M., Bogaert, M., & Benoit, D. (2025). Evaluating the stability of model explanations in instance dependent cost-sensitive credit scoring. Joint ORBEL-NGB Conference https://biblio.ugent.be/publication/01JS4JWMWY3J2QWQC0R0N5ZBZ7 [21] Calzarossa, M.C., Giudici, P. and Zieni, R. (2025). An assessment framework for explainable AI with applications to cybersecurity, Artificial Intelligence Review 58 (150) https://link.springer.com/article/10.1007/s10462-025-11141-w
[22] Hammond, L., et al. (2025). Multi-Agent Risks from Advanced AI (Technical report#1). Cooperative AI Foundation. https://arxiv.org/abs/2502.14143v1
[23] Stiefenhofer, P. (2025), Artificial General Intelligence and the End of Human Employment: The Need to Renegotiate the Social Contract https://arxiv.org/abs/2502.07050v1
[24] Occhipinti, J., et al. (2024). In the Shadow of Smith's Invisible Hand: Risks to Economic Stability and Social Wellbeing in the Age of Intelligence https://arxiv.org/abs/2407.01545v1
[25] Occhipinti, J., et al. (2024). Recessionary Pressures of Generative AI: A Threat to Wellbeing. https://arxiv.org/abs/2403.17405v1
[26] Kulveit, J., et al. (2025), Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development. https://arxiv.org/abs/2501.16946v2
[27] International Monetary Fund (2024), Broadening the Gains from Generative AI: The Role of Fiscal Policies. https://www.imf.org/-/media/Files/Publications/SDN/2024/English/SDNEA2024002.ashx
[28] Knowles, B., D'Cruz, J., Richards, J.T., and Varshney, K. R. (2023). "Humble AI." Communications of the ACM 66, no. 9: 73-79 https://dl.acm.org/doi/pdf/10.1145/3587035
[29] WIRED Magazine (June 2023). Humans Aren't Mentally Ready for an AI-Saturated 'Post-Truth World' https://www.wired.com/story/generative-aideepfakes-disinformation-psychology/
[30] Gefen, D., et al. (2025). The Importance of Distrust in Trusting Digital Worker Chatbots. Communications of the ACM. https://cacm.acm.org/research/the-importance-of-distrust-in-trusting-digital-worker-chatbots/
[31] Cohen, M., et al. (2024). Believing Anthropomorphism: Examining the Role of Anthropomorphic Cues on Trust in Large Language Models. Proceedings of the ACM

Conference on Human Factors in Computing Systems (CHI). https://dl.acm.org/doi/10.1145/3613905.3650818

[32] Beavins, E. (2025, April). CHAI embarks on post-deployment monitoring for AI as FDA lags. Fierce Healthcare. https://www.fiercehealthcare.com/ai-and-machine-learning/chai-embarks-post-deployment-monitoring-ai