

Más allá del prompt: los vacíos de la IA agéntica en la justicia colombiana

Alberto Mario Zambrano Romero¹

El artículo analiza la distinción entre la IA generativa o reactiva —de una sola iteración— de la IA agéntica, que planifica y ejecuta autónomamente cadenas de microacciones con mínima supervisión humana. A partir del Acuerdo PCSJA24-12243 (2024) del Consejo Superior de la Judicatura, diseñado para el formato *prompt*-respuesta, se identifican casos de uso que permiten evidenciar vacíos frente a la utilización de IA agéntica en la administración de justicia. El artículo contrasta estos vacíos con el Reglamento (UE) 2024/1689 y concluye con los retos pendientes para el ordenamiento colombiano.

SUMARIO:

I. Introducción. Tipos de IA: IA generativa o reactiva vs. IA agéntica; II. Casos diferenciados según niveles de riesgo; III. Los cinco vacíos de la IA agéntica en el Acuerdo PCSJA24-12243; IV. ¿Qué contempla el Reglamento (UE) 2024/1689 que resulta adaptable y complementario al Acuerdo PCSJA24-12243?; V. Conclusiones; VI. Bibliografía.

I. Introducción. Tipos de IA: generativa o reactiva vs. agéntica

El IBM Data and AI Team en 2023 definió cuatro diferentes tipos de IA:

¹ Alberto es abogado de la Universidad del Norte, especialista en Propiedad Intelectual, Derecho de Autor y Nuevas Tecnologías de la Universidad Externado de Colombia y magíster en Propiedad Intelectual de la Universidad Carlos III de Madrid. Cuenta con siete años de experiencia de trabajo en firmas como OlarteMoure & Asociados en el área de Innovación, Protección y Transferencia (IP&T), como Gestor de PI en la Dirección de Innovación de la Pontificia Universidad Javeriana, asesor jurídico de empresas de software tipo SaaS como Lexica S.A.S. y actualmente como Director de Legal Ops en Ariel.

- a) IA mecánica reactiva,
- b) IA reactiva o de memoria limitada,
- c) teoría de la mente en IA,
- d) IA autoconsciente².

La IA **mecánica reactiva**³ está diseñada para llevar a cabo una tarea muy específica y no puede recordar resultados o decisiones anteriores ya que solo trabaja con datos existentes en el presente. Es usualmente utilizada para analizar grandes volúmenes de datos. El ejemplo más nombrado de la IA mecánica reactiva es IBM Deep Blue —desarrollado por IBM—, conocido por haber vencido al famoso campeón mundial de ajedrez Garry Kasparov.

La IA **reactiva**, denominada también como IA **de memoria limitada**⁴, puede recordar eventos y resultados pasados, monitorizar objetos o situaciones a lo largo del tiempo y utilizar datos del pasado y el presente para decidir su acción próxima. Estos datos no se retienen, pero sí se utilizan para mejorar su rendimiento o entrenar su modelo. Aquí es donde encasillan a las herramientas de IA generativa, asistentes virtuales y *chatbots* como Siri, Alexa, ChatGPT, entre otros.

Ahora bien, respecto de los últimos dos tipos (**teoría de la mente en IA** e **IA autoconsciente**) es importante mencionar que, si bien a la fecha no han sido desarrolladas, sí se prevé que estas podrán comprender los pensamientos y emociones de otras entidades. Esto permitiría simular relaciones similares a las humanas y responder mejor a estados emocionales humanos en los que la IA autoconsciente sería capaz de comprender sus propias condiciones y rasgos internos.

Es así como, a partir del auge tecnológico de los años recientes, las personas se han venido familiarizando en su día a día con uno que otro modelo. Posiblemente, la más conocida a la fecha de escritura de esta columna es la IA de memoria limitada o, para efectos del presente artículo, IA generativa o reactiva. Aquella que opera bajo el formato *prompt* → respuesta, en donde el usuario formula una instrucción, el sistema genera una salida en una única iteración y el humano decide qué

² IBM Data and AI Team. «Comprender los diferentes tipos de inteligencia artificial». IBM. 12 de octubre de 2023. <https://www.ibm.com/es-es/think/topics/artificial-intelligence-types>.

³ Arend Hintze, «Understanding the Four Types of AI, from Reactive Robots to Self-Aware Beings». *The Conversation*, 14 de noviembre de 2016.

⁴ OECD. «The Agentic AI Landscape and Its Conceptual Foundations». *OECD Artificial Intelligence Papers*, n.º 56. París: OECD Publishing, 2026.

hacer con ella. Cada acto de generación es verificable y atribuible a una instrucción concreta dada por el usuario.

De hecho, este es el modelo contemplado en la definición de «sistemas o herramientas de inteligencia artificial generativa» del Acuerdo PCSJA24-12243, que describe un sistema que genera contenido «en respuesta a instrucciones en interfaces conversacionales de lenguaje natural (*prompts*)», fiel a la tecnología que dominaba el mercado cuando el acuerdo fue expedido: el *chatbot* conversacional de propósito general.

Sin embargo, hoy en día ya existe la IA **agéntica** que opera bajo un formato distinto: un objetivo de alto nivel → planificación autónoma⁵ → ejecución de una cadena de microacciones → resultado consolidado. El sistema no se limita a responder: descompone el objetivo en subtarear, decide el orden de ejecución, invoca herramientas externas⁶ y encadena esas acciones con una supervisión humana que puede ser previa, posterior o, en los diseños más autónomos, prácticamente inexistente. Por ello, este artículo se pregunta si el Acuerdo PCSJA24-12243 ampara el uso de IA agéntica en la administración de justicia colombiana y, según el nivel de riesgo, qué elementos del Reglamento (UE) 2024/1689 podrían complementarlo.

Para responder esas preguntas primero hay que analizar dos fenómenos técnicos propios del formato de IA agéntica que merecen nombre desde ahora, porque reaparecen a lo largo de este análisis. El primero es el riesgo acumulativo y compuesto: si cada microacción de una cadena tiene una probabilidad individual de error, la posibilidad de que al menos una de estas numerosas microacciones esté viciada crece sustancialmente, y el error de un paso intermedio puede propagarse y contaminar silenciosamente los pasos siguientes⁷. El segundo es la deriva de objetivo (*goal drift*)⁸: la posibilidad de que el agente, en su planificación autónoma, reinterprete el objetivo para el cual se creó originalmente de un modo que técnicamente lo cumple, pero se aparta de la intención razonable de quien lo formuló.

⁵ Pattie Maes, «Artificial Life Meets Entertainment: Lifelike Autonomous Agents», *Communications of the ACM* 38, n.º 11 (1995): 108–114.

⁶ Shunyu Yao *et al.*, «ReAct: Synergizing Reasoning and Acting in Language Models». Presentado en la 11.ª International Conference on Learning Representations (ICLR 2023), Kigali, Ruanda. <https://arxiv.org/abs/2210.03629>.

⁷ Anushree Sinha *et al.*, «Human-on-the-Loop Orchestration for AI-Assisted Legal Discovery», 2026, arXiv:2606.19812. <https://arxiv.org/abs/2606.19812>.

⁸ Rauno Arike *et al.*, «Technical Report: Evaluating Goal Drift in Language Model Agents», 2025, arXiv:2505.02709. <https://arxiv.org/abs/2505.02709>.

II. Casos diferenciados según niveles de riesgo

Ahora bien, tres casos, ordenados según el grado de independencia de la IA y la complejidad de las tareas que realiza, ilustran los diferentes niveles de riesgo de la IA agéntica en la administración de justicia, y sirven de hilo conductor al resto del artículo:

- **Caso 1 — Agente de investigación jurisprudencial (riesgo bajo).** Recibe un tema de estudio y devuelve una síntesis de líneas jurisprudenciales relevantes con sus fuentes. Ejecuta autónomamente varias búsquedas, filtra resultados y consolida un informe, pero su producto es un insumo de consulta que el servidor judicial revisa íntegramente antes de usar: el punto de verificación humana posterior cubre razonablemente el riesgo; sin embargo, no elimina la posibilidad de que se produzcan alucinaciones de fuentes según la herramienta.
- **Caso 2 — Agente de gestión procesal (riesgo medio-alto).** Extrae y clasifica información de expedientes, identifica procesos repetitivos, apoya su asignación y puede redactar documentos de forma autónoma: una cadena que pasa, sin pausa ni autorización intermedia, de tareas puramente administrativas a tareas que rozan la motivación de una decisión judicial. La verificación humana posterior sigue existiendo, el servidor revisa el proyecto antes de firmarlo, pero ya no cubre los eslabones intermedios.
- **Caso 3 — Agente de optimización de descongestión judicial (riesgo alto).** Recibe el objetivo de «maximizar la evacuación de procesos represados» y decide autónomamente el orden de asignación, prioriza unos expedientes sobre otros, genera alertas de reparto y actualiza el estado de los procesos, orientado por una métrica de rendimiento o un sistema de recompensa que desconoce el servidor judicial. Aquí el riesgo ya no es solo el error puntual: es la posibilidad de que la propia función de optimización, correctamente ejecutada, produzca resultados sistemáticamente lesivos debido al *goal-drift* (como postergar indefinidamente los procesos complejos porque no brindan las mismas métricas de evacuación).

En el próximo acápite analizaremos si los anteriores casos de uso resultan aplicables en la administración de justicia, de la mano del contenido del Acuerdo PCSJA24-12243 y el Reglamento (UE) 2024/1689 frente a la IA agéntica.

III. Los cinco vacíos de la IA agéntica en el Acuerdo PCSJA24-12243 y el Reglamento (UE) 2024/1689

3.1. LOS CINCO VACÍOS DE LA IA AGÉNTICA EN REFERENCIA AL ACUERDO PCSJA24-12243

En Colombia, el acuerdo construyó un ciclo de gobernanza razonablemente completo y, para su momento, diciembre de 2024, fue pionero en Iberoamérica. Fue expedido, en alianza con la UNESCO, en cumplimiento del exhorto de la Corte Constitucional en la Sentencia T-323 de 2024⁹, y sobre la evidencia de una encuesta interna en la que cerca del 30 % de los servidores ya usaba herramientas de IA y el 91,68 % consideraba necesaria la adopción de lineamientos¹⁰.

El acuerdo está compuesto, *grosso modo*, por: 1) principios y niveles de uso; 2) deberes y control; y 3) implementación institucional. Al comparar el contenido del acuerdo con los tres casos graduados del capítulo anterior, según su nivel de independencia, se advierte que fue pensado para un escenario en el que un servidor judicial formula un *prompt*, recibe una respuesta y decide, caso por caso, si usa la respuesta. Este enfoque revela los cinco vacíos que dejó el acuerdo frente a la IA agéntica, y que se desarrollan a continuación:

Primer vacío: ausencia de una definición de «agente» y «acción autónoma». El acuerdo define seis categorías construidas sobre el formato *prompt*-respuesta, pero ninguna cubre a una IA agéntica. Un despacho que active el caso 2 no encontraría en el acuerdo un supuesto de hecho claro al cual asimilarlo. La definición de *prompt* del mismo artículo agrava el problema, porque asume una correspondencia uno a uno entre instrucción y resultado: en el caso 2, un único objetivo de alto nivel desencadena una decena de resultados intermedios que el servidor judicial nunca instruyó individualmente. Ese nexo no es justamente lo que el artículo 10⁽¹¹⁾ exige documentar cuando ordena hacer explícitos en el expediente «los *prompts* o instrucciones utilizadas».

⁹ C. Const. , Sent. T-323, ago. 2/2024. M. P. Juan Carlos Cortés González.

¹⁰ Consejo Superior de la Judicatura, Acuerdo PCSJA24-12243, 2024.

¹¹ *Ibid.*, art. 10.

Segundo vacío: falta de trazabilidad de los pasos intermedios. El artículo 10 exige documentar en el expediente el nombre de la herramienta, modelo y empresa proveedora, la fecha de uso, la aceptación de los términos y condiciones, la finalidad del uso, los *prompts* utilizados y la ubicación específica de los productos incorporados. Es un régimen de transparencia serio para el formato reactivo: documenta la entrada y la salida. Sin embargo, el acuerdo no exige ni prevé mecanismo alguno para registrar la cadena de microacciones intermedias que un agente ejecuta entre la instrucción inicial y el resultado consolidado. En el caso 2, si el proyecto de auto contiene un error, reconstruir en qué subtarea de la cadena se originó resulta problemático pero imprescindible.

Esa reconstrucción tiene, en primer lugar, una finalidad probatoria: permitir que el servidor judicial pueda demostrar que ejerció la verificación diligente que exige el artículo 9.2⁽¹²⁾ consistente en «contrastar y verificar la veracidad, relevancia y suficiencia de los resultados obtenidos», y que la autoridad disciplinaria pueda establecer dónde se desvió la cadena. En segundo lugar, tiene una finalidad conceptual: el artículo 2.2⁽¹³⁾ define tres riesgos (alucinaciones, desinformación y sesgo de IA) predicados de un resultado único frente a una entrada única.

El acuerdo previó el riesgo del contenido resultante, no el riesgo del proceso que lo genera, y por eso el riesgo acumulativo y compuesto descrito en el capítulo II (el error de un paso en la cadena que contamina los siguientes) no tiene categoría en la taxonomía del acuerdo, pese a ser el riesgo característico del formato agéntico.

Tercer vacío: indefinición sobre dónde y cómo deben ubicarse los hitos de supervisión humana. El artículo 4 escalona los usos permitidos en tres niveles según su proximidad a la función jurisdiccional:

- **Nivel 1 (art. 4.1):** Gestión administrativa y de soporte.
- **Nivel 2 (art. 4.2):** Usos sujetos a revisión detallada.
- **Nivel 3 (art. 4.3):** Simulación de escenarios de decisión y asistencia en tareas que inciden en la motivación de las decisiones judiciales.

La escala gradúa el riesgo, pero no el control y, por ello, el acuerdo no indica en qué momento de cada nivel debe intervenir el servidor judicial ni cómo documentar esa intervención. Esa gradación

¹² *Ibid.*, art. 9.2.

¹³ *Ibid.*, art. 2.2.

es razonable en su lógica interna: a mayor cercanía con el núcleo de la función jurisdiccional, mayor exigencia.

El problema es que un solo flujo agéntico puede atravesar los tres niveles en una sola ejecución: el agente del caso 2 transita de extraer información de expedientes (4.2.g) a clasificar procesos repetitivos (4.2.j) y a redactar un auto que, si roza la motivación de una decisión sustancial, entra en el terreno del 4.3.c. El acuerdo no dice qué ocurre cuando una tarea autónoma cruza ese umbral, dado que solo regula el tipo de uso, no la transición entre *tipos de uso* dentro de una misma ejecución. El principio que podría resolver esta problemática es el art. 3.9, que exige «realizar un estricto escrutinio sobre las actuaciones y decisiones en las cuales usen herramientas de IA».

Cuarto vacío: ausencia de un esquema de responsabilidad escalonada. En el plano de los deberes individuales, el artículo 8 impone al servidor judicial, entre otros, evitar *chatbots* comerciales gratuitos (8.1), prescindir de herramientas opacas en el origen de sus datos para los usos 4.2 y 4.3 (8.2), prescindir de IA para la valoración probatoria, el escrutinio de hechos y la solución de problemas jurídicos (8.3) y, críticamente, «abstenerse de introducir, por cualquier medio, datos personales, información sensible o confidencial» cuando la herramienta los reutilice (8.6). Todos esos deberes se formulan en primera persona, sobre la premisa de un acto voluntario y consciente del servidor.

En el caso 3, es el propio agente quien decide, en ejecución autónoma, qué datos de un expediente trasladar entre sistemas para optimizar el reparto, sin que exista un acto humano equivalente al que el numeral 8.6 del acuerdo sanciona. Si el agente traslada el dato sensible, ¿responde el servidor bajo un estándar de responsabilidad por el hecho propio, por culpa *in vigilando*, o estamos ante un vacío que exige una categoría nueva?

El Capítulo 3 (arts. 11-12) traslada la gobernanza a la Unidad de Transformación Digital e Informática (UTDI) frente al análisis de impacto algorítmico, la evaluación de la transparencia de los datos de entrenamiento, la seguridad de la información, la prohibición de usar datos personales en el entrenamiento, el análisis de impacto de privacidad y la priorización de proveedores auditables cuyos términos y condiciones sean compatibles con la normativa nacional, pero regula la adopción con una evaluación única y *ex ante*, realizada al momento de habilitarla, no sobre la ejecución continuada. El resultado: si el agente del caso 3 prioriza mal un expediente y se deriva un perjuicio, el acuerdo no

ofrece una regla que determine si responde el servidor que autorizó el objetivo, la UTDI que avaló el despliegue o el proveedor.

Quinto vacío: falta de validación de los sistemas de recompensa. El caso 3 lo expone con mayor nitidez: un agente de optimización de descongestión actúa orientado por una métrica (evacuar el mayor número de procesos) que ninguna disposición del acuerdo exige auditar, precisamente porque no consideró el formato agéntico. No obstante, un sistema de recompensa mal calibrado podría incentivar al agente a priorizar la velocidad de evacuación sobre la calidad o la garantía de derechos. El artículo 11.4⁽¹⁴⁾ exige «procesos propios de calidad de los datos [...] con el objeto de disminuir sesgos y alucinaciones», y se acerca al problema, pero no lo nombra.

A estos cinco vacíos normativos se suma una variable adicional que el acuerdo no diagnostica y que condiciona la eficacia de cualquiera de sus reglas: una parte significativa de las relatorías digitales de los tribunales superiores del distrito en Colombia continúa en formato físico o en repositorios no digitalizados e indexados. Esto ha significado un reto en curso desde hace varios años en donde la estabilización de las tareas de digitalización de los expedientes se esperaba lograr en algunos meses siguientes a la expedición del documento de la Rama Judicial «Seguimiento al Plan Estratégico de Transformación Digital» de septiembre de 2021⁽¹⁵⁾.

Los usos de búsqueda y síntesis jurisprudencial del artículo 4.2⁽¹⁶⁾ (el corazón del caso 1) dependen de un universo digital razonablemente completo: si una porción sustancial de la jurisprudencia relevante permanece en archivo físico, el agente completa el patrón esperado con la información parcial que sí tiene, y produce una síntesis que aparenta ser completa sin serlo; el riesgo de alucinación del artículo 2.2 se materializa no por invención, sino por vacío documental que el modelo no puede advertir.

La trazabilidad hacia adentro del expediente que exige el artículo 10 presupone una trazabilidad hacia afuera, que se rompe cuando esa fuente no está digitalizada. Un deber de transparencia que se apoya en una infraestructura desigual entre despachos termina exigiéndose con mayor rigor donde ya existe digitalización, y diluyéndose donde el riesgo de opacidad es mayor.

¹⁴ *Ibid.*, art. 11.4.

¹⁵ Consejo Superior de la Judicatura, Seguimiento al Plan Estratégico de Transformación Digital (Bogotá: Rama Judicial de Colombia, 2021), 22.

¹⁶ art. 4.2. Consejo Superior de la Judicatura, Acuerdo PCSJA24-12243, 2024.

Cerrar estos vacíos no es solo tarea normativa, es también una tarea de capacitación. Por ello, el artículo 13⁽¹⁷⁾ encomienda a la Escuela Judicial Rodrigo Lara Bonilla, en articulación con la UTDI, a incorporar una ruta de competencias digitales y de IA, con «posibilidad de alianzas con la academia, centros de pensamiento, investigación e innovación». Esa cláusula es, en realidad, la condición de eficacia de todo lo anterior: enseñar a supervisar una cadena autónoma, leer un registro de trazabilidad, distinguir una derivación de objetivo de un error puntual y reconocer el momento en que la cadena cruzó un umbral de riesgo exigen una comprensión de la arquitectura del sistema que excede las capacidades tecnológicas presentes de la rama judicial.

En síntesis, los cinco vacíos comparten una causa común: el acuerdo construyó sus reglas sobre la unidad de análisis del *prompt* y su respuesta en un momento en el que la tecnología dominante era mayoritariamente reactiva. No son fallas de redacción corregibles con un ajuste semántico, son la consecuencia lógica de haber regulado el *prompt* y no la cadena.

3.2. LOS CINCO VACÍOS DE LA IA AGÉNTICA EN REFERENCIA AL REGLAMENTO (UE) 2024/1689

El instrumento europeo objeto de análisis es el Reglamento (UE) 2024/1689⁽¹⁸⁾ del Parlamento Europeo y del Consejo, del 13 de junio de 2024, conocido como el «AI Act». Un reglamento es de aplicación directa y obligatoria en toda la Unión, sin necesidad de transposición nacional, lo que explica su nivel de detalle operativo (plazos, umbrales, listas de obligaciones) frente a la naturaleza más principal del acuerdo colombiano, y lo convierte en el referente comparado más útil para medir los cinco vacíos.

El artículo 1 fija como objeto promover una IA «centrada en el ser humano y fiable», y el artículo 2⁽¹⁹⁾ delimita el ámbito de aplicación por actores de la cadena de valor (proveedores, responsables del despliegue, importadores, distribuidores, con alcance extraterritorial). Esa arquitectura basada en actores, y no solo en usos o tipos de usuario final como la del artículo 1 del acuerdo, es la que permite después distribuir responsabilidad según la posición de cada interviniente.

¹⁷ *Ibid.*, art. 13.

¹⁸ Reglamento (UE) 2024/1689.

¹⁹ *Ibid.*, arts. 1 y 2.

Sobre el primer vacío: ausencia de una definición de «agente» y «acción autónoma». El artículo 3.1⁽²⁰⁾ define «sistema de IA» como «un sistema basado en una máquina diseñado para funcionar con distintos niveles de autonomía y que puede mostrar capacidad de adaptación tras el despliegue [...] para objetivos explícitos o implícitos». A partir de esa definición, sostengo que la IA agéntica puede quedar comprendida dentro de la definición existente, para cubrir así un sistema que planifica subtarefas y adapta su comportamiento sobre la marcha.

Sobre el segundo vacío: falta de trazabilidad de los pasos intermedios. El artículo 12 «exige que los sistemas de alto riesgo permitan técnicamente el registro automático de acontecimientos [...] a lo largo de todo el ciclo de vida del sistema». El artículo 19 obliga al proveedor a conservar esos registros o *logs* por al menos seis meses, y el artículo 26.6 impone la misma obligación simétrica al responsable del despliegue. El artículo 13 sobre transparencia y comunicación de información exige que el sistema vaya acompañado de instrucciones de uso con un contenido mínimo tasado, incluidas las medidas de supervisión humana del artículo 14 y «una descripción de los mecanismos [...] que permita [...] recabar, almacenar e interpretar correctamente los archivos de registro» (13.3.f)²¹.

Frente al artículo 10 del acuerdo, que documenta la entrada y la salida, el sistema europeo documenta el proceso minuciosamente: un registro automático y continuo de acontecimientos, exactamente la unidad de análisis que la cadena del caso 2 requeriría para reconstruir en qué subtarea se originó un error.

Sobre el tercer vacío: indefinición sobre dónde y cómo deben ubicarse los hitos de supervisión humana. El artículo 14 es la pieza de mayor relevancia, pues no formula un principio de conducta como el 3.9 del acuerdo, sino una especificación técnica exigible al diseño del sistema. Su apartado 1⁽²²⁾ exige que los sistemas de alto riesgo «se diseñarán [...] de modo que puedan ser vigilados de manera efectiva por personas físicas», dotados de interfaces humano-máquina adecuadas: no basta con que el operador decida supervisar, el sistema debe estar construido para permitirlo.

Su apartado 3 exige que las medidas sean «proporcionales a los riesgos, al nivel de autonomía y al contexto de uso», y constituye la única referencia expresa, en todo el corpus normativo examinado

²⁰ *Ibid.*, art. 3.1.

²¹ *Ibid.*, art. 12, 13, 13.3., 14, 19 y 26.6.

²² *Ibid.*, art. 14.1.

en este artículo, a la graduación de la intensidad de la supervisión según la variable que el agente introduce. Su apartado 4 detalla lo que el supervisor debe poder hacer: a) comprender las capacidades y limitaciones; b) ser consciente del sesgo de automatización; c) interpretar correctamente los resultados; d) decidir, en cualquier situación concreta, no utilizar el sistema o descartar sus resultados; y e) intervenir en su funcionamiento o interrumpirlo. El apartado 5 añade, para los sistemas más sensibles, una regla de verificación por al menos dos personas físicas independientes.

Sobre el cuarto vacío: ausencia de un esquema de responsabilidad escalonada. Los artículos 6, 8, 16, 17, 25 y 26 componen la arquitectura de imputación distribuida que el acuerdo no tiene. El artículo 6 clasifica el alto riesgo por reglas objetivas (incluyendo los sistemas para la administración de justicia), con una válvula de escape para el proveedor que considere que su sistema no plantea riesgo importante (6.4). Operan sobre esa clasificación la gestión de riesgos como «proceso iterativo continuo» (art. 9), las doce obligaciones del proveedor del artículo 16 y el sistema de calidad del artículo 17, que incluye «un marco de rendición de cuentas que defina las responsabilidades del personal» (17.1.m)²³.

Los artículos 25 y 26⁽²⁴⁾ responden la pregunta que el acuerdo deja sin resolver: cuando algo sale mal, ¿quién responde y en qué momento? La lógica es una escalera de tres peldaños según quién tiene el sistema en cada fase: mientras el sistema es diseñado y puesto en el mercado, el proveedor responde por la conformidad, calidad y documentación (arts.16 y 17). Una vez la herramienta está en uso, la responsabilidad pasa al responsable del despliegue, que en nuestro contexto sería el despacho judicial: debe poner al frente a personas con competencia y autoridad para supervisarla (26.2); si detecta indicios de riesgo, debe dejar de usarlo de inmediato y avisar al proveedor (26.5); debe guardar los registros durante al menos seis meses (26.6), igual que el proveedor (art. 19); y, si es autoridad pública y usa los sistemas más sensibles, necesita autorización previa y debe rendir informes anuales (26.10). Pero si alguien altera el sistema, la responsabilidad se traslada a él. Quien diseña responde por el diseño, quien opera responde por la vigilancia y la suspensión oportuna, y quien altera asume el lugar de quien diseñó.

²³ *Ibid.*, art. 17.1.m.

²⁴ *Ibid.*, arts. 25 y 26.

Sobre el quinto vacío: falta de validación de los sistemas de recompensa. El reglamento no ofrece una respuesta específica. El artículo 9 exige que la gestión de riesgos sea un proceso continuo durante todo el ciclo de vida, y el artículo 15.4 exige que los sistemas que «continúan aprendiendo tras su introducción en el mercado» se diseñen para eliminar o reducir el riesgo de que se generen resultados sesgados. Ambas se aproximan al escenario del caso 3, en donde un agente, cuya función de optimización orienta decisiones sucesivas, aprende y se retroalimenta de ellas; sin embargo, ninguna exige auditar de forma autónoma y previa la función objetivo o el mecanismo de recompensa que determina cómo el agente prioriza sus microacciones. Este es, de los cinco vacíos, el único que persiste como problema abierto en ambos ordenamientos.

IV. ¿Qué contempla el Reglamento (UE) 2024/1689 que resulta adaptable y complementario al Acuerdo PCSJA24-12243?

Ambos instrumentos comparten la misma matriz de valores (primacía de la decisión humana, graduación del riesgo, transparencia, prevención de sesgos), y la coincidencia no es casual: el propio acuerdo declara en sus considerandos haber tomado como referente los instrumentos de UNESCO, OCDE y CEPEJ, la misma familia de principios que inspira al reglamento. La diferencia no es de valores, sino de técnica de implementación: donde el acuerdo proclama principios y confía su aplicación a la diligencia individual del servidor judicial, el reglamento traduce esos mismos principios en obligaciones documentales, especificaciones técnicas de diseño y reglas de imputación verificables por una autoridad de vigilancia del mercado europeo.

Una ruta de formación específicamente agéntica debería organizarse, en consecuencia, por fase de la cadena y por responsable de esa fase, y no solo por perfil de cargo. Según la fase de la cadena agéntica, el contenido formativo mínimo podrá variar. En instancias iniciales el servidor judicial o el despacho deberá definir el objetivo del agente, para que así el proveedor tecnológico pueda, durante la etapa de planificación autónoma, reflejar el plan construido en conjunto y, antes de su ejecución, validar con el despacho el sistema de recompensas. Posteriormente, y durante la ejecución de las microacciones, tanto el servidor como el sistema deben tener conocimiento y preparación para

analizar el registro de *logs* para que en determinado caso aquellas actividades en donde se detecte una anomalía puedan ser pausadas y revisadas por un humano.

Así mismo, de la comparación surge una agenda concreta y acotada para una eventual reforma del Acuerdo PCSJA24-12243, ordenada según los cinco vacíos:

1. Redefinir el «sistema de IA» mediante la incorporación expresa de los niveles de autonomía y la capacidad de adaptación tras el despliegue, siguiendo el artículo 3.1⁽²⁵⁾ del reglamento, de modo que la definición cubra a los agentes, sin forzar su interpretación extensivamente. Complementarla, además, con las nociones de «cadena de acciones» y «acción autónoma».
2. Exigir que la IA lleve un registro automático de eventos «logs» a lo largo del ciclo de ejecución de sus actividades como agente, con período mínimo de conservación y obligaciones simétricas para el proveedor y para el despacho usuario, en concordancia con los artículos 12, 19 y 26.6⁽²⁶⁾.
3. Traducir la supervisión humana en especificaciones técnicas exigibles, interfaz de vigilancia, facultad de descartar o revertir resultados y un mecanismo de interrupción segura equivalente al botón de parada del artículo 14.4.e⁽²⁷⁾, graduadas según el nivel de autonomía del sistema.
4. Incorporar una regla de responsabilidad escalonada inspirada en los artículos 16, 17, 25 y 26: obligaciones diferenciadas por actor, un marco de rendición de cuentas exigible, una regla de conversión de estatus ante modificaciones sustanciales o derivas de objetivo, y un deber activo de suspensión inmediata del sistema ante indicios de riesgo, en reemplazo del deber de reporte del artículo 14.3.
5. Asumir la validación de los sistemas de recompensa como frente propio de regulación: exigir que, antes de habilitar cualquier agente de optimización, ya sea de reparto, de descongestión o de priorización, se documente y audite de forma independiente la función objetivo contra la cual el sistema fue entrenado u optimizado, con la verificación correspondiente de su compatibilidad con las garantías procesales.

²⁵ *Ibid.*, art. 3.1.

²⁶ *Ibid.*, arts. 12, 19 y 26.6.

²⁷ *Ibid.*, 14.4.e.

Esto sugiere que, si bien el marco europeo ofrece respuestas sólidas para la definición conceptual del agente, la exigencia de trazabilidad de su actividad, el rol de la supervisión humana y la responsabilidad escalonada de los diferentes actores no terminan de proporcionar un modelo adaptable frente a la validación técnica de los sistemas de recompensa que orientan el comportamiento autónomo de los agentes.

V. Conclusiones

Primero, la distinción entre la IA reactiva o generativa resulta determinante para valorar el riesgo jurídico de su uso en la administración de justicia, pues introduce dos fenómenos ausentes en el formato reactivo: el riesgo acumulativo y compuesto propio de las cadenas de microacciones y la deriva de objetivo (*goal drift*). Esta distinción se confirma en los tres casos de uso graduados examinados: el agente de investigación jurisprudencial (riesgo bajo), el agente de gestión procesal (riesgo medio-alto) y el agente de optimización de descongestión judicial (riesgo alto), que muestran que el nivel de riesgo no depende únicamente de la tarea asignada al agente, sino del grado de independencia con que la ejecuta y de si existen, o no, puntos de verificación humana efectivos a lo largo de toda la cadena y no solo sobre su resultado final.

Segundo, el Acuerdo PCSJA24-12243 acierta en sus principios, pero regula el *prompt* y el resultado, no la cadena de decisiones intermedias propia de la IA agéntica. La aplicación hipotética de los tres casos de uso graduados permite concluir que los cinco vacíos no son fallas de redacción, sino la consecuencia de haber sido diseñado en un momento en que la tecnología dominante era reactiva. A ellos se suma una precondition estructural que ningún régimen normativo puede suplir: la digitalización incompleta del acervo jurisprudencial.

Tercero, el Reglamento (UE) 2024/1689 aporta un conjunto de elementos técnicamente adaptables y complementarios al Acuerdo PCSJA24-12243 frente a cuatro de los cinco vacíos identificados: (i) una definición del sistema de IA construida sobre los niveles de autonomía y la capacidad de adaptación tras el despliegue, que podría incorporarse junto con las nociones de «cadena de acciones» y «acción autónoma»; (ii) un régimen de registro automático y continuo de eventos (*logs*), con obligaciones simétricas entre proveedor y desplegador, y un período mínimo de conservación; (iii) la traducción de la supervisión humana en especificaciones técnicas exigibles al diseño del sistema y no solo en un principio de conducta, graduadas según el nivel de autonomía y con un

mecanismo de interrupción segura equivalente al botón de parada; y (iv) una arquitectura de responsabilidad escalonada y distribuida según el actor y el control efectivo sobre cada fase de la cadena de valor, con regla de conversión de estatus ante modificaciones sustanciales o derivas de objetivo.

Estos cuatro elementos, sin embargo, no agotan la respuesta: la validación de los sistemas de recompensa que orientan la optimización autónoma del agente permanece como un problema normativo abierto tanto en el ordenamiento colombiano como en el europeo, lo que exige para ese quinto vacío una solución propia que ninguno de los dos instrumentos ofrece todavía. En síntesis, el artículo 25⁽²⁸⁾ del reglamento europeo ofrece el andamiaje conceptual más cercano a esa figura; adaptarlo a la administración de justicia colombiana podría suplir ese vacío.

Cuarto, en respuesta a la pregunta problema, el uso de la IA agéntica en la administración de justicia colombiana no se encuentra amparado de manera uniforme por el Acuerdo PCSJA24-12243: lo está, con suficiencia razonable, en supuestos de riesgo bajo como el caso 1, donde la verificación humana posterior sobre un producto de consulta cubre adecuadamente el riesgo, pero deja de estarlo a medida que la autonomía del agente crece, porque el acuerdo fue construido sobre el formato de IA reactiva, no sobre la cadena de decisiones intermedias que caracteriza a los casos 2 y 3 de IA agéntica. Los cinco vacíos identificados no son, por tanto, fallas de redacción corregibles puntualmente, sino la consecuencia de haber regulado un formato tecnológico distinto al que hoy plantea la realidad.

El Reglamento (UE) 2024/1689, como referente de derecho comparado más desarrollado, ofrece una respuesta técnica suficiente para cuatro de esos cinco vacíos con elementos plenamente adaptables al ordenamiento colombiano sin necesidad de reproducir su nivel de detalle operativo. El quinto vacío, la validación de los sistemas de recompensa, permanece como un problema abierto en ambos ordenamientos y exige, más allá del ejercicio comparado, una solución normativa propia.

En suma, la IA agéntica no está prohibida ni plenamente amparada por el acuerdo vigente, está insuficientemente regulada, y su uso responsable en la rama judicial dependería más de una actualización, inspirada quizás, mas no limitada, en el reglamento europeo de manera que traduzca

²⁸ *Ibid.*, art. 25.

los principios ya declarados por el acuerdo en obligaciones verificables a lo largo de toda la cadena agéntica, y no solo sobre su resultado final.

VI. Bibliografía

- Arike, Rauno, Elizabeth Donoway, Henning Bartsch, y Marius Hobbhahn. «Technical Report: Evaluating Goal Drift in Language Model Agents», 2025, arXiv:2505.02709. <https://arxiv.org/abs/2505.02709>.
- Hintze, Arend. «Understanding the Four Types of AI, from Reactive Robots to Self-Aware Beings». *The Conversation*, 14 de noviembre de 2016. <https://theconversation.com/understanding-the-four-types-of-ai-from-reactive-robots-to-self-aware-beings-67616>.
- IBM Data and AI Team. «Tipos de inteligencia artificial». *IBM*, 12 de octubre de 2023. <https://www.ibm.com/es-es/think/topics/artificial-intelligence-types>.
- Maes, Pattie. «Artificial Life Meets Entertainment: Lifelike Autonomous Agents». *Communications of the ACM*, 38, n.º 11 (1995): 108–114.
- OECD. «The Agentic AI Landscape and Its Conceptual Foundations». *OECD Artificial Intelligence Papers*, n.º 56. París: OECD Publishing, 2026. <https://doi.org/10.1787/396cf758-en>.
- Sinha, Anushree, Srivaths Ranganathan, Abhishek Dharmaratnakar y Debanshu Das. «Human-on-the-Loop Orchestration for AI-Assisted Legal Discovery», 2026. arXiv:2606.19812. <https://arxiv.org/abs/2606.19812>.
- Yao, Shunyu, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan y Yuan Cao. «ReAct: Synergizing Reasoning and Acting in Language Models» Presentado en la 11.ª International Conference on Learning Representations (ICLR 2023), Kigali, Ruanda. <https://arxiv.org/abs/2210.03629>.

Jurisprudencia y normativa – Colombia:

Consejo Superior de la Judicatura. *Seguimiento al Plan Estratégico de Transformación Digital*. Bogotá: Rama Judicial de Colombia, 2021.

Consejo Superior de la Judicatura de Colombia. *Acuerdo PCSJA24-12243, "Por el cual se adoptan lineamientos para el uso y aprovechamiento respetuoso, responsable, seguro y ético de la inteligencia artificial en la Rama Judicial"* (16 de diciembre de 2024).

Corte Constitucional de Colombia. Sentencia T-323 del 2 de agosto de 2024. M. P. Juan Carlos Cortés González.

Normativa internacional

Parlamento Europeo y Consejo de la Unión Europea. «Reglamento (UE) 2024/1689, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (Reglamento de Inteligencia Artificial)». *Diario Oficial de la Unión Europea*, 12 de julio de 2024.

El análisis de los artículos 1, 2, 3, 6, 8, 10, 13, 14, 16, 17, 19, 25 y 26 del Reglamento (UE) 2024/1689 se realizó sobre el texto oficial en español publicado en el Diario Oficial de la Unión Europea (DO L de 12 de julio de 2024).