De la garantía contractual a la garantía arquitectónica: MencIA y la soberanía del diseño ante la inteligencia artificial generativa en el sector público local. Una lectura del RGPD en su décimo aniversario

From contractual to architectural safeguards: MencIA and sovereignty by design facing generative artificial intelligence in the local public sector. A reading of the GDPR at its tenth anniversary

José Antonio Arias Lombardero
Jefe de la Sección de Innovación y Participación Ciudadana. Diputación Provincial de Lugo
ORCID: https://orcid.org/0009-0007-7772-1008

Fecha de recepción: 20.06.2026
Fecha de aceptación: 01.08.2026
Fecha de publicación en línea: octubre 2026

Resumen:
La adopción de inteligencia artificial generativa en las administraciones locales tensiona el modelo de garantías sobre el que descansa el Reglamento General de Protección de Datos, de naturaleza esencialmente contractual: opacidad del entrenamiento, auditabilidad limitada, extraterritorialidad estructural y polivalencia funcional desbordan los presupuestos de tecnologías acotadas, finalidades determinables y encargados identificables. Este trabajo propone la noción de soberanía arquitectónica como traducción técnica del cumplimiento desde el diseño del artículo 25 RGPD y la examina mediante un estudio de caso: MencIA, un ecosistema de inteligencia artificial en código abierto desplegado sobre infraestructura propia por la Diputación de Lugo. El análisis del artefacto, contrastado artículo por artículo con el Reglamento y con la política de la Agencia Española de Protección de Datos sobre IA

Abstract:
The adoption of generative artificial intelligence in local administrations strains the safeguards model underlying the General Data Protection Regulation, which is essentially contractual in nature: training opacity, limited auditability, structural extraterritoriality and functional polyvalence exceed its assumptions of bounded technologies, determinable purposes and identifiable processors. This paper proposes the notion of architectural sovereignty as the technical translation of data protection by design under Article 25 GDPR and examines it through a case study: MencIA, an open-source artificial intelligence ecosystem deployed on proprietary infrastructure by the Provincial Council of Lugo. The analysis of the artefact, tested article by article against the Regulation and against the Spanish Data Protection Agency's policy on generative AI (November 2025), shows that local execution simplifies

[1] Declaración de uso de inteligencia artificial. En la elaboración de este manuscrito se ha empleado inteligencia artificial generativa (Claude 4.8 Opus y 4.6 Sonnet, Anthropic), como herramienta de apoyo para la estructuración del texto, la revisión gramatical y la revisión de normas de citación. El contenido sustantivo, los argumentos, el análisis jurídico y las conclusiones son responsabilidad exclusiva del autor, que ha revisado, validado y supervisado críticamente la totalidad del texto.

Declaración sobre fuentes de financiación, apoyos institucionales y posibles conflictos de interés. Esta investigación no ha recibido financiación externa ni apoyo institucional para su realización, y no responde a encargo alguno. La autoría de este trabajo ha participado en el diseño del sistema que se analiza, un proyecto de la Diputación de Lugo desarrollado con medios de esa entidad, y figura entre los firmantes de dos de las referencias bibliográficas citadas, en las que el modelo se formuló con anterioridad a su adopción institucional. Ambas circunstancias deben ponderarse al valorar las afirmaciones descriptivas sobre el funcionamiento del sistema, que se sostienen en la inspección directa de su código y de su configuración, según se precisa en el apartado 5.1. La autoría no percibe contraprestación alguna, directa o indirecta, vinculada al sistema examinado ni a las tecnologías empleadas en él.

generativa (noviembre de 2025), muestra que la ejecución local simplifica el análisis de responsabilidad, sostiene la trazabilidad y la repetibilidad de las operaciones y mantiene la información sensible dentro del perímetro de la administración. Como prueba de validez prospectiva, el modelo neutraliza por diseño las configuraciones de riesgo que las orientaciones de la Agencia sobre IA agéntica (febrero de 2026) consideran inaceptables, sin perjuicio de las limitaciones propias de una prueba de concepto operativa.

Palabras clave:

Soberanía arquitectónica; inteligencia artificial generativa; protección de datos desde el diseño; código abierto; administración local.

Sumario:

1. Introducción. **1.1.** El Reglamento General de Protección de Datos en su décimo aniversario. **1.2.** La irrupción de la inteligencia artificial generativa en el sector público local. **1.3.** Tesis y estructura del trabajo. **2.** Marco normativo y doctrinal. **2.1.** El régimen del Reglamento General de Protección de Datos. **2.2.** El Reglamento de Inteligencia Artificial y la supervisión humana. **2.3.** La política de la Agencia Española de Protección de Datos sobre IA generativa. **2.4.** Las orientaciones sobre inteligencia artificial agéntica. **2.5.** La articulación entre regulación y *soft law*. **3.** La fatiga del modelo contractual ante la inteligencia artificial generativa. **3.1.** Los presupuestos arquitectónicos implícitos del RGPD. **3.2.** Cuatro tensiones que introduce la inteligencia artificial generativa. **3.3.** Los límites de los contratos de encargo ante el artículo 9. **4.** La soberanía arquitectónica como garantía. **4.1.** Delimitación del concepto. **4.2.** La política de la Agencia como bisagra doctrinal. **4.3.** Los componentes operativos. **5.** Estudio de caso: el ecosistema MencIA. **5.1.** Caracterización del artefacto. **5.2.** Cumplimiento desde el diseño, artículo por artículo. **5.3.** Una prueba de validez prospectiva: la arquitectura ante la inteligencia artificial agéntica. **5.4.** Las limitaciones de la fase inicial. **6.** Discusión y limitaciones. **6.1.** El riesgo de absolutizar la soberanía arquitectónica. **6.2.** Dependencias residuales. **6.3.** Auditabilidad necesaria, no suficiente. **6.4.** Una línea prospectiva: la inteligencia artificial como tecnología de mejora de la privacidad. **7.** Conclusiones. Bibliografía.

the responsibility assessment, sustains the traceability and repeatability of operations and keeps sensitive information within the administration's perimeter. As a prospective validity test, the model neutralises by design the risk configurations that the Agency's guidance on agentic AI (February 2026) deems unacceptable, subject to the limitations of an operational proof of concept.

Keywords:

Architectural sovereignty; generative artificial intelligence; data protection by design; open source; local government.

Summary:

***1.** Introduction. **1.1.** The General Data Protection Regulation at its tenth anniversary. **1.2.** The emergence of generative artificial intelligence in the local public sector. **1.3.** Thesis and structure of the paper. **2.** Regulatory and doctrinal framework. **2.1.** The regime of the General Data Protection Regulation. **2.2.** The Artificial Intelligence Regulation and human oversight. **2.3.** The policy of the Spanish Data Protection Agency on generative AI. **2.4.** The guidance on agentic artificial intelligence. **2.5.** The articulation between regulation and soft law. **3.** The fatigue of the contractual model facing generative artificial intelligence. **3.1.** The implicit architectural assumptions of the Regulation. **3.2.** Four tensions introduced by generative artificial intelligence. **3.3.** The limits of data processing agreements under Article 9. **4.** Architectural sovereignty as a safeguard. **4.1.** Delimiting the concept. **4.2.** The Agency's policy as a doctrinal hinge. **4.3.** The operational components. **5.** Case study: the MencIA ecosystem. **5.1.** Characterisation of the artefact. **5.2.** Compliance by design, article by article. **5.3.** A prospective validity test: the architecture facing agentic artificial intelligence. **5.4.** Limitations of the initial phase. **6.** Discussion and limitations. **6.1.** The risk of absolutising architectural sovereignty. **6.2.** Residual dependencies. **6.3.** Auditability as necessary, not sufficient. **6.4.** A prospective line: artificial intelligence as a privacy-enhancing technology. **7.** Conclusions. Bibliography.*

1. Introducción

1.1 El Reglamento General de Protección de Datos en su décimo aniversario

El Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos (en adelante, RGPD), cumple diez años desde su adopción habiendo consolidado un modelo de garantías que descansa, en buena medida, sobre instrumentos de naturaleza contractual y documental: el consentimiento como acto de voluntad, los contratos de encargo del artículo 28, las cláusulas contractuales tipo para transferencias internacionales y las evaluaciones de impacto como ejercicio previo de diligencia. Ese andamiaje fue diseñado para un ecosistema tecnológico en el que las finalidades del tratamiento podían determinarse con precisión, los encargados eran identificables y trazables, y la lógica de los sistemas resultaba, al menos en principio, inspeccionable. La eficacia del modelo dependía de un supuesto implícito rara vez explicitado: que la garantía jurídica podía construirse sobre la palabra dada y verificarse sobre el papel firmado.

La administración local española ha operado durante una década dentro de ese marco con un grado razonable de seguridad jurídica. Las entidades locales tratan datos de categorías especiales en ámbitos como los servicios sociales, la recaudación o la gestión de personal, y han venido apoyando su cumplimiento en la combinación de contratos de encargo con proveedores tecnológicos y en la confianza depositada en infraestructuras de terceros. Esa confianza, sin embargo, presupone que el proveedor cumple lo pactado, que el dato permanece donde se acordó y que la finalidad declarada agota los usos efectivos de la información.

1.2 La irrupción de la inteligencia artificial generativa en el sector público local

La adopción de inteligencia artificial generativa altera ese equilibrio. A diferencia del software tradicional, los grandes modelos de lenguaje se entrenan sobre corpus cuya composición resulta opaca, operan con una polivalencia funcional que difícilmente se deja reconducir al principio de finalidad y, cuando se consumen como servicio en la nube, exponen la información a cadenas de procesamiento sujetas a legislaciones extraterritoriales. La tensión no es meramente teórica: una herramienta que procesa correos electrónicos con datos de salud o de situaciones de violencia de género, o que sintetiza expedientes de servicios sociales, plantea de inmediato la cuestión de dónde se procesa el dato, quién puede acceder a él y con qué finalidad ulterior puede ser reutilizado.

Frente a este escenario, las administraciones disponen de dos respuestas arquetípicas. La primera consiste en reforzar el modelo contractual existente —adendas de protección de datos más detalladas, garantías reforzadas, auditorías de proveedor— manteniendo la dependencia de infraestructuras externas. La segunda consiste en desplazar la garantía desde la cláusula hacia la propia ingeniería del sistema, internalizando la infraestructura de procesamiento de modo que el dato no abandone el perímetro de la administración. La Agencia Española de Protección de Datos se ha enfrentado a esta disyuntiva en su propia organización y la ha resuelto en favor de la segunda vía: su política general interna para el uso de IA generativa en procesos administrativos, de noviembre de 2025, establece que los casos de uso que traten información personal o sensible se implementen preferentemente en sistemas internos o ad hoc, y reserva los sistemas externos para supuestos acotados con garantías reforzadas[2]. El documento posee alcance autoorganizativo y no constituye interpretación oficial dirigida al conjunto del sector público. Su interés para este análisis, como se precisará más adelante, reside precisamente en que es la autoridad de control quien, al decidir sobre sus propios tratamientos, opta por la arquitectura antes que por el contrato.

1.3 Tesis y estructura del trabajo

Este trabajo examina la segunda respuesta a través de un estudio de caso. MencIA, la estrategia de inteligencia artificial de la Diputación de Lugo, constituye un ecosistema de herramientas y asistentes construido íntegramente sobre modelos de código abierto y pesos abiertos, ejecutados en infraestructura propia. El argumento que se defiende es que MencIA materializa lo que aquí se denomina soberanía arquitectónica: la traducción técnica del principio de protección de datos desde el diseño y por defecto del artículo 25 del RGPD, según la cual determinadas garantías que el modelo contractual confía a la voluntad de las partes pasan a quedar incorporadas como propiedades estructurales del sistema, no susceptibles de incumplimiento por una decisión posterior.

[2] Agencia Española de Protección de Datos, *Política general interna para el uso de IA generativa* en procesos administrativos de la AEPD (27 de noviembre de 2025).

La aportación se ofrece con una calibración expresa. MencIA se encuentra en una fase inicial de laboratorio y los resultados que se presentan tienen el valor de una prueba de concepto operativa en entorno real, no el de un estudio de impacto a escala provincial. Esta naturaleza preliminar, lejos de debilitar el análisis, delimita con honestidad su alcance y previene una sobrevaloración de las conclusiones.

El trabajo se estructura en cinco partes tras esta introducción. La segunda reconstruye el marco normativo y doctrinal aplicable. La tercera analiza las cuatro tensiones que la inteligencia artificial generativa introduce en el modelo contractual del RGPD. La cuarta propone y delimita el concepto de soberanía arquitectónica. La quinta examina el caso MencIA, contrastándolo artículo por artículo con el RGPD e incorporando una prueba de validez prospectiva frente a las orientaciones de la Agencia sobre inteligencia artificial agéntica de febrero de 2026. La sexta discute los límites del modelo y las dependencias que subsisten, antes de las conclusiones.

2. Marco normativo y doctrinal

2.1 El régimen del Reglamento General de Protección de Datos

El análisis se ancla en un conjunto acotado de preceptos del RGPD. El artículo 5 enuncia los principios de licitud, limitación de la finalidad, minimización, exactitud, limitación del plazo de conservación e integridad y confidencialidad, que operan como criterio de validez de cualquier tratamiento. El artículo 9 prohíbe, con excepciones tasadas, el tratamiento de categorías especiales de datos, un precepto de especial relevancia para la administración local, que maneja de forma ordinaria información sobre salud, situación socioeconómica o circunstancias personales sensibles. El artículo 22 reconoce el derecho a no ser objeto de decisiones basadas únicamente en el tratamiento automatizado que produzcan efectos jurídicos o afecten significativamente al interesado.

Tres preceptos adicionales vertebran el argumento. El artículo 25 consagra la protección de datos desde el diseño y por defecto, exigiendo que el responsable aplique medidas técnicas y organizativas apropiadas tanto en el momento de determinar los medios del tratamiento como durante su ejecución. El artículo 28 regula la figura del encargado del tratamiento y el contenido mínimo del contrato que ha de vincularle con el responsable. Los artículos 44 y siguientes someten las transferencias internacionales a un régimen de garantías cuya finalidad es evitar que el nivel de protección reconocido en la Unión se vea menoscabado al exportar el dato. La articulación de estos preceptos define el modelo de garantías cuya suficiencia ante la inteligencia artificial generativa se examina más adelante.

2.2 El Reglamento de Inteligencia Artificial y la supervisión humana

El Reglamento (UE) 2024/1689, de Inteligencia Artificial (en adelante, RIA), incorpora una pieza directamente conectada con el modelo que aquí se analiza. Su artículo 14 impone la supervisión humana sobre los sistemas de inteligencia artificial de alto riesgo, exigiendo que las personas físicas encargadas puedan comprender el sistema, vigilar su funcionamiento e intervenir sobre él o interrumpirlo. La supervisión humana funciona de este modo como obligación de diseño del operador: la capacidad de procesamiento del sistema no desplaza la autoridad resolutiva ni la responsabilidad, que permanecen en una persona identificada.

Este mandato opera en un plano distinto, aunque complementario, al del artículo 22 del RGPD. Mientras el RIA impone a quien despliega el sistema la obligación de diseñarlo con supervisión humana efectiva, el artículo 22 del RGPD reconoce al ciudadano afectado el derecho a obtener intervención humana por parte del responsable cuando una decisión automatizada produzca efectos jurídicos en él o le afecte significativamente. El primero es una obligación de arquitectura dirigida al operador; el segundo es un derecho subjetivo del interesado frente al responsable del tratamiento. Ambos convergen en el mismo principio de administración aumentada: la máquina propone y el empleado público dispone, pero desde planos normativos distintos que el diseño del sistema debe satisfacer simultáneamente.

2.3 La política de la Agencia Española de Protección de Datos sobre IA generativa

En el plano del soft law, la política general interna para el uso de IA generativa en procesos administrativos publicada por la Agencia Española de Protección de Datos en noviembre de 2025, acompañada de su anexo de implementación, ofrece un criterio de particular interés. El documento clasifica los

sistemas de inteligencia artificial generativa en función de su grado de control por la organización y establece una preferencia jerarquizada: los casos de uso que impliquen tratamiento de información personal o sensible deben implementarse preferentemente en sistemas internos o ad hoc, esto es, desplegados bajo control de la propia entidad, reservando los sistemas externos para supuestos en los que existan garantías reforzadas y no se traten datos de alto riesgo.[3]

Conviene delimitar con precisión el alcance de este instrumento. Se trata de una política de naturaleza autoorganizativa, que la propia Agencia circunscribe a sus procesos administrativos internos y de la que excluye expresamente todo efecto de verificación, evaluación o aplicación indirecta del RIA, así como cualquier pretensión de interpretación oficial dirigida al conjunto del sector público. Fuera de la propia Agencia, por tanto, su fuerza no es la de una recomendación con destinatario general, sino la de un soft law persuasivo cuyo valor deriva de su autoría. Esta delimitación no afecta al núcleo del argumento que aquí se desarrolla, que se apoya de forma principal en las orientaciones sobre inteligencia artificial agéntica examinadas en el apartado siguiente, publicadas estas sí con vocación de dirigirse al sector.

El interés de esta política para el presente análisis reside, así precisado su alcance, en el desplazamiento que documenta. La autoridad que supervisa el cumplimiento del RGPD, al decidir sobre sus propios tratamientos, no resuelve el problema reforzando las garantías contractuales que podría exigir a un proveedor, sino eligiendo dónde y bajo qué control se ejecuta el sistema. La conformidad deja de descansar únicamente sobre la calidad de las cláusulas pactadas con un tercero y pasa a depender de una decisión de arquitectura. Que ese criterio provenga de un instrumento autoorganizativo lo priva de fuerza prescriptiva general, pero conserva su valor como precedente cualificado sobre cómo la autoridad de control entiende que se materializa la protección desde el diseño cuando el tratamiento emplea esta tecnología.

2.4 Las orientaciones sobre inteligencia artificial agéntica

Esta misma orientación se proyecta sobre escenarios de mayor autonomía. En febrero de 2026 la Agencia publicó unas orientaciones sobre inteligencia artificial agéntica desde la perspectiva de protección de datos, en las que advierte que la ejecución autónoma y encadenada de tareas multiplica los intervinientes y dificulta la trazabilidad de los flujos de datos.[4] No se invocan aquí como objeto del trabajo, sino como criterio externo para comprobar, en el estudio de caso, si una arquitectura concebida para la IA generativa conserva su capacidad de garantía frente a tecnologías más autónomas.

2.5 La articulación entre regulación y *soft law*

El marco descrito combina normas vinculantes —el RGPD y el RIA— con instrumentos de soft law emanados de la autoridad de control. La doctrina ha advertido que esta combinación, característica de la gobernanza tecnológica europea, traslada al responsable una carga interpretativa considerable, en la medida en que las orientaciones de las autoridades concretan principios abiertos sin agotar su contenido.[5] La política de la Agencia sobre IA generativa y sus orientaciones sobre IA agéntica no crean obligaciones nuevas, sino que precisan cómo cumplir las existentes en un contexto tecnológico inédito. Esta función de concreción es la que permite leer dichos documentos como bisagra entre el mandato abstracto del artículo 25 del RGPD y las decisiones técnicas concretas que un responsable debe adoptar, tránsito que la noción de soberanía arquitectónica pretende articular conceptualmente.

[3] Agencia Española de Protección de Datos, *Política general interna …,* y Anexo implementación de la Política general IAG de la AEPD (11 de diciembre de 2025).

[4] Agencia Española de Protección de Datos, *Inteligencia artificial agéntica desde la perspectiva de protección de datos*, v1.2 (febrero de 2026).

[5] Sobre la función concretadora del *soft law* de las autoridades de protección de datos en contextos de innovación tecnológica, véase Lorenzo Cotino Hueso, "Qué concreta transparencia e información de algoritmos e inteligencia artificial es la debida", *Revista Española de la Transparencia*, n.º 16 (2023): 17-63.

3. La fatiga del modelo contractual ante la inteligencia artificial generativa

3.1 Los presupuestos arquitectónicos implícitos del RGPD

El modelo de garantías reconstruido en el apartado anterior reposa sobre una premisa que rara vez se enuncia porque, durante años, se cumplió sin esfuerzo: que la garantía jurídica puede edificarse sobre vínculos contractuales entre partes identificables. El consentimiento vincula al interesado con el responsable; el contrato del artículo 28 vincula al responsable con el encargado; las cláusulas contractuales tipo extienden esa red de obligaciones hasta el importador de datos en un tercer país. En todos los casos, la protección del dato se confía a una obligación asumida por un sujeto determinado, cuyo cumplimiento se presume y, llegado el caso, se exige.

Este diseño funciona bien cuando concurren tres condiciones de fondo. La primera es que las finalidades del tratamiento puedan determinarse de antemano con suficiente precisión, de modo que el principio de limitación de la finalidad del artículo 5 opere como criterio aplicable. La segunda es que los encargados que intervienen en la cadena de tratamiento sean identificables y su actuación trazable, requisito sin el cual el contrato del artículo 28 carece de objeto verificable. La tercera es que la lógica del sistema empleado resulte, al menos en principio, inspeccionable, de manera que la diligencia del responsable pueda proyectarse sobre el funcionamiento real del tratamiento y no solo sobre su descripción documental. Cuando estas tres condiciones se cumplen, el modelo contractual del RGPD despliega su eficacia con normalidad.

3.2 Cuatro tensiones que introduce la inteligencia artificial generativa

La inteligencia artificial generativa comercial, consumida como servicio en la nube, tensiona las tres condiciones anteriores en cuatro puntos que conviene examinar por separado.

La primera tensión es la opacidad del entrenamiento. Los grandes modelos de lenguaje se construyen sobre corpus cuya composición exacta no se hace pública y cuyo proceso de entrenamiento el responsable del tratamiento no puede inspeccionar ni reproducir. Esta opacidad no es un defecto coyuntural subsanable mediante una cláusula, sino una característica estructural del producto: el responsable contrata un sistema cuya génesis le resulta inaccesible, lo que erosiona la posibilidad de proyectar sobre él una diligencia informada. La doctrina ha señalado que esta opacidad desplaza la protección de datos desde el terreno de la verificación hacia el de la confianza, precisamente el terreno que el RGPD pretendía abandonar.[6]

La segunda tensión es la auditabilidad limitada. Aun cuando el responsable quisiera verificar qué datos entran efectivamente en el sistema y cómo se procesan, la arquitectura propietaria de los servicios comerciales lo impide. El responsable puede auditar el contrato, pero no el sistema; puede comprobar que existe una cláusula que prohíbe la reutilización de los datos para entrenamiento, pero no puede verificar técnicamente que esa prohibición se respeta. Se produce así una disociación entre el plano documental, donde el cumplimiento parece asegurado, y el plano técnico, donde permanece incomprobable.

La tercera tensión es la extraterritorialidad estructural. Cuando el procesamiento se realiza en infraestructuras de proveedores sujetos a legislaciones de terceros países, el dato queda potencialmente expuesto a marcos normativos que pueden imponer accesos al margen de las garantías del RGPD. La legislación estadounidense en materia de acceso a datos por las autoridades ilustra el problema: las cláusulas contractuales tipo, concebidas para garantizar un nivel de protección equivalente, pueden resultar inoperantes frente a un mandato legal extraterritorial que obligue al proveedor con independencia de lo pactado. La jurisprudencia del Tribunal de Justicia de la Unión Europea en materia de transferencias internacionales ha evidenciado reiteradamente la fragilidad de las garantías meramente contractuales cuando el ordenamiento del país de destino no ofrece protección equivalente.[7]

La cuarta tensión es la polivalencia funcional. A diferencia del software especializado, un modelo generativo es por naturaleza polivalente: el mismo sistema puede redactar, resumir, clasificar, traducir

[6] Sobre la opacidad algorítmica y las limitaciones de la diligencia del responsable ante sistemas propietarios, véase Michael Veale e Irina Brass, "Administration by Algorithm? Public Management Meets Public Sector Machine Learning", en *Algorithmic Regulation*, ed. por Karen Yeung y Martin Lodge (Oxford: Oxford University Press, 2019), 121-149.

[7] Sentencias del Tribunal de Justicia de la Unión Europea en los asuntos C-362/14, *Maximillian Schrems c. Data Protection Commissioner* (Schrems I), de 6 de octubre de 2015, y C-311/18, *Data Protection Commissioner c. Facebook Ireland Ltd y Maximillian Schrems* (Schrems II), de 16 de julio de 2020.

o inferir sobre datos de la más diversa índole. Esta versatilidad, que constituye su principal atractivo, resiste el principio de limitación de la finalidad, pues el sistema no está diseñado para un tratamiento determinado sino para una indeterminación de tratamientos posibles. La finalidad deja de ser una propiedad del sistema para convertirse en una restricción que ha de imponérsele desde fuera, mediante decisiones de configuración que el modelo contractual no contempla.

3.3 Los límites de los contratos de encargo ante el artículo 9

Estas cuatro tensiones se agudizan cuando el tratamiento recae sobre categorías especiales de datos. La administración local trata de forma ordinaria información de salud, situación socioeconómica o circunstancias de especial vulnerabilidad, protegida por el artículo 9 del RGPD. En este ámbito, la insuficiencia del modelo contractual se vuelve crítica: un contrato de encargo, por exhaustivo que sea, no impide técnicamente que el dato sensible transite por una infraestructura de terceros, no garantiza por sí mismo que no sea reutilizado y no neutraliza la exposición a legislaciones extraterritoriales. El responsable que confía la protección de datos de violencia de género o de expedientes de servicios sociales exclusivamente a la robustez de una adenda contractual asume un riesgo cuya materialización no puede verificar ni, llegado el caso, revertir.

Conviene precisar el alcance de esta crítica. No se sostiene aquí que el modelo contractual del RGPD sea inválido o haya quedado derogado por la irrupción de la inteligencia artificial generativa. El régimen del artículo 28 y de las transferencias internacionales conserva plena vigencia y resulta idóneo para numerosos tratamientos. Lo que se afirma es más limitado: que ese modelo, suficiente para tecnologías acotadas y finalidades determinables, muestra signos de fatiga cuando se proyecta sobre sistemas opacos, inauditables, extraterritoriales y polivalentes que tratan datos de categorías especiales. Esa fatiga no reclama la sustitución del modelo contractual, sino su complemento por garantías de otra naturaleza, incorporadas a la propia arquitectura del sistema. La articulación de ese complemento es el objeto del apartado siguiente.

4. La soberanía arquitectónica como garantía

4.1 Delimitación del concepto

La sección anterior ha identificado cuatro tensiones que la inteligencia artificial generativa introduce en el modelo contractual del RGPD. Conviene preguntarse si esas tensiones admiten una respuesta dentro del propio modelo, o si reclaman un complemento de distinta naturaleza. La tesis de este trabajo es que ese complemento existe y puede formularse con precisión: consiste en trasladar a la arquitectura del sistema garantías que el modelo contractual confía a la voluntad de las partes.

La expresión soberanía arquitectónica designa aquí un principio de diseño según el cual el responsable del tratamiento incorpora las garantías de protección de datos como propiedades estructurales del sistema de información, de modo que determinados incumplimientos dejan de ser conductas que un tercero puede cometer para convertirse en estados que el sistema no puede alcanzar. La diferencia con la protección por contrato es funcional, no jerárquica: no se sostiene que la arquitectura sustituya al contrato, sino que lo complementa en aquellos puntos donde la garantía contractual resulta inverificable, como se ha expuesto respecto de la opacidad del entrenamiento y la auditabilidad limitada de los sistemas propietarios.

El concepto no es original en sus componentes: la protección de datos desde el diseño del artículo 25, la minimización por defecto y la seguridad del artículo 32 apuntan en la misma dirección. Lo que el concepto añade es la articulación de esos mandatos como decisión de infraestructura, no como política organizativa. Cuando el artículo 25 exige medidas técnicas apropiadas en el momento de determinar los medios del tratamiento, la soberanía arquitectónica responde que la elección de ejecutar los modelos en infraestructura propia, con código inspeccionable y sobre corpus cerrados, es esa medida técnica, y que su eficacia no depende de que alguien la cumpla, sino de que el sistema está construido así.

4.2 La política de la Agencia como bisagra doctrinal

La política general interna para el uso de IA generativa en procesos administrativos publicada por la Agencia Española de Protección de Datos en noviembre de 2025 opera, con el alcance autoorganizativo ya delimitado, como bisagra entre el mandato abstracto del artículo 25 y la decisión concreta de arquitectura. Al clasificar los sistemas de inteligencia artificial generativa en función de su grado

de control por la organización y al establecer que los casos de uso que traten información personal o sensible se implementen preferentemente en sistemas internos o ad hoc, la Agencia no enuncia una obligación nueva ni la impone a terceros: resuelve, para sus propios tratamientos, dónde entiende que debe materializarse la protección desde el diseño cuando el tratamiento emplea esta tecnología. Esa decisión interna es la que aquí se lee como indicio cualificado de que el perímetro del dato debe coincidir con el perímetro de la responsabilidad.

Este desplazamiento tiene una consecuencia que la propia Agencia explicita en un documento posterior. En sus orientaciones sobre inteligencia artificial agéntica de febrero de 2026, al analizar la determinación de responsabilidades de tratamiento, la Agencia formula una asimetría que merece transcribirse: «Cuando un agente de IA se ejecute de forma totalmente local no cabría más análisis de la responsabilidad».[8] La formulación es relevante porque no se limita a reconocer que la ejecución local reduce el riesgo; afirma que cierra el análisis de responsabilidad, en el sentido de que no genera relaciones con encargados, subencargados ni destinatarios adicionales que sea necesario evaluar. Cualquier otra configuración, por el contrario, abre ese análisis y obliga a recorrer la cadena completa de intervinientes.

La asimetría resultante puede enunciarse como regla: a medida que el procesamiento se externaliza, las obligaciones de cumplimiento se multiplican; a medida que se internaliza, se simplifican. No se trata de que el procesamiento externo sea ilícito —el RGPD lo permite con garantías—, sino de que cada capa de externalización añade un estrato de obligaciones cuyo cumplimiento depende de la conducta de un tercero que el responsable no controla directamente. La soberanía arquitectónica invierte esa lógica: en lugar de gestionar capas sucesivas de riesgo delegado, elimina la delegación en la capa donde reside la información sensible.

4.3 Los componentes operativos

Formulado el concepto y anclado en la doctrina de la autoridad de control, conviene descomponer la soberanía arquitectónica en los elementos que la hacen operativa. El análisis de la literatura existente sobre el caso y de los documentos fundacionales del artefacto que se examina en la sección siguiente permite identificar cuatro componentes.

El primero es la ejecución local de los modelos de inferencia. Los modelos de lenguaje se ejecutan en servidores con procesamiento gráfico dedicado ubicados en las dependencias de la administración, de modo que los prompts, los documentos analizados y las salidas generadas no transitan por infraestructuras de terceros. Este componente neutraliza directamente la tercera tensión identificada, la extraterritorialidad estructural, y simplifica el análisis de responsabilidad en los términos que la Agencia formula.

El segundo es la posibilidad de inspección del código. La utilización de modelos de código abierto y de pesos abiertos, y de un conjunto tecnológico íntegramente auditable, permite que la función interventora, la persona delegada de protección de datos y, llegado el caso, la autoridad de supervisión examinen el funcionamiento real del sistema. La segunda tensión, la auditabilidad limitada, se aborda en su raíz: la disociación entre el plano documental y el plano técnico desaparece cuando el código que ejecuta el tratamiento puede leerse.

El tercero es el confinamiento del conocimiento mediante técnicas de recuperación aumentada sobre corpus cerrados. Cada asistente accede únicamente a la base documental que tiene asignada, depurada y validada por la propia institución, sin acceso a internet ni a fuentes externas no controladas. Este componente opera simultáneamente sobre la primera tensión, porque limita la opacidad al acotar el dominio de conocimiento del modelo, y sobre la cuarta, porque transforma un sistema polivalente en un sistema de finalidad acotada por diseño.

El cuarto es la contención de la autonomía. Las herramientas operan bajo un paradigma de tarea única en el que la máquina propone y el empleado público dispone: carecen de permisos lógicos para ejecutar acciones con efectos jurídicos, de modo que la supervisión humana no es una directriz que deba observarse, sino un límite técnico que no puede franquearse. Este componente conecta con el artículo 22 del RGPD, así como con el artículo 14 del RIA y adquiere un relieve particular cuando se proyecta sobre escenarios de inteligencia artificial agéntica, como se examinará en el estudio de caso.

Estos cuatro componentes no operan de forma aislada. Su eficacia reside en su articulación: la ejecución local pierde parte de su valor si el código es opaco; el código abierto es condición necesaria

[8] Agencia Española de Protección de Datos, *Inteligencia artificial agéntica*, v1.2, apartado V.A, "Determinación de responsabilidades de tratamiento", 33.

pero no suficiente si el modelo accede a fuentes no controladas; y el confinamiento del conocimiento requiere, para traducirse en garantía de protección de datos, que la autonomía del sistema esté contenida de modo que sus salidas no generen efectos irreversibles sin intervención humana. La soberanía arquitectónica es, en este sentido, una propiedad del conjunto, no de sus partes.

5. Estudio de caso: el ecosistema MencIA

5.1 Caracterización del artefacto

MencIA es la estrategia de inteligencia artificial de la Diputación de Lugo, una administración provincial del noroeste de España que presta servicios a sesenta y siete municipios, de los cuales uno solo supera los veinte mil habitantes. El componente que aquí se analiza es su laboratorio de inteligencia artificial, un conjunto de herramientas y asistentes en funcionamiento que materializa, en un entorno administrativo real, los principios de la soberanía arquitectónica.[9] El marco estratégico que ampara el laboratorio fue aprobado por unanimidad por el Pleno de la Diputación el 26 de mayo de 2026.

Desde la perspectiva metodológica, el artefacto se examina como un estudio de caso de tipo prueba de concepto. No se pretende medir su impacto a escala provincial ni evaluar resultados consolidados de una implantación madura, sino mostrar la factibilidad técnica y organizativa de un modelo de garantías incorporadas a la arquitectura. Esta delimitación es deliberada: el laboratorio se encuentra en una fase inicial y sus resultados tienen valor demostrativo, no estadístico. La evaluación adopta un enfoque cualitativo y dogmático, centrado en el contraste de la arquitectura del sistema con las exigencias normativas del RGPD. La fase actual excluye deliberadamente la medición de tiempos de inferencia, volumen de uso o tasas de precisión estadística, al priorizar la validación jurídica de la conformidad por diseño antes de abordar el escalado operativo.

La base de conocimiento sobre la que se sostienen las afirmaciones relativas al comportamiento del sistema debe explicitarse. Las propiedades arquitectónicas que se describen —ejecución local de la inferencia, acotación del corpus mediante recuperación aumentada y ausencia de capacidad de actuación sobre los sistemas de registro— se han verificado mediante inspección directa del código fuente y de la configuración efectiva del despliegue, y no se infieren de la documentación estratégica. Quien firma este trabajo ha participado en el desarrollo del artefacto, circunstancia que sitúa el análisis en la posición metodológica propia de la investigación en diseño: el acceso privilegiado a la configuración del sistema permite afirmar sus propiedades con precisión, pero impide presentarlas como verificación independiente. El apartado sexto retoma esta limitación al examinar el alcance real de la auditabilidad.

La arquitectura se articula en tres capas, todas ejecutadas en infraestructura propia. La capa de inferencia ejecuta modelos de lenguaje de código abierto y pesos abiertos mediante un gestor local, sobre servidores con procesamiento gráfico dedicado alojados en las dependencias de la administración. La capa de aplicaciones, contenida en entornos aislados, ofrece tanto una interfaz general de conversación como un conjunto de aplicaciones especializadas de tarea única. La capa de orquestación conecta las interfaces con los modelos y con bases documentales propias mediante técnicas de generación aumentada por recuperación, garantizando que cada asistente acceda únicamente al corpus que tiene asignado y no a fuentes externas. Sobre esta arquitectura operan más de una quincena de herramientas y asistentes, entre los que figuran un gestor de correo electrónico, un asistente de contratación pública que opera sobre un corpus normativo cerrado, una calculadora de plazos administrativos, un monitor del boletín oficial para el servicio de recaudación y diversos asistentes especializados en ámbitos concretos de la gestión. No todas las piezas del catálogo emplean inferencia: algunas, como el monitor del boletín oficial, resuelven su cometido mediante recuperación y filtrado documental sin intervención de un modelo de lenguaje, lo que reduce la superficie de riesgo asociada al tratamiento mediante modelos generativos.

El rasgo que define a este conjunto es lo que la propia estrategia denomina el paradigma de tarea única: cada herramienta resuelve un problema acotado con un flujo de interacción mínimo. Conviene

[9] El modelo que la estrategia analizada acabó adoptando —infraestructura propia, modelos de código abierto y pesos abiertos, y proyección supramunicipal desde el nivel provincial— se formuló en sus líneas esenciales con anterioridad en Arias Lombardero, "Gobierno local inteligente y soberano: una propuesta estratégica para la implementación de IA de código abierto en la provincia de Lugo", *Revista digital CEMCI*, n.º 67 (2025), y en Rivera Capón, Arias Lombardero y Fariña Verea, "Gobernanza algorítmica local: diseño e implementación de un *stack* de IA soberana en código abierto para entidades supramunicipales", *IDP. Revista de Internet, Derecho y Política*, n.º 43 (2025), ambos anteriores a la aprobación institucional de la estrategia. El presente trabajo se distingue de ellos en su objeto: no propone el modelo ni describe su arquitectura, sino que examina su capacidad para satisfacer las exigencias del RGPD mediante garantías incorporadas al diseño. Sobre la relación de estas referencias con la autoría de este trabajo, véase la declaración que cierra el artículo.

precisar, sin embargo, que el ecosistema comprende dos tipos de piezas con comportamientos distintos. Las aplicaciones son herramientas de consulta única y salida estructurada: el usuario aporta un material, la herramienta produce un resultado concreto en una sola interacción. Los asistentes especializados son por naturaleza conversacionales y admiten varias interacciones por sesión, aunque siempre sobre corpus cerrados y dentro de un dominio normativo acotado. Lo que garantiza el cumplimiento de protección de datos en ambos casos no es la mínima interacción, sino el dominio acotado y la ejecución local: ninguna pieza, sea aplicación o asistente, externaliza el dato ni accede a fuentes ajenas al corpus que tiene asignado.

Esta distinción obliga a responder una objeción que la propia descripción suscita: si el carácter conversacional permite varias interacciones por sesión, cabe preguntar si no reintroduce la polivalencia funcional identificada como cuarta tensión. La respuesta exige distinguir entre variedad de formulación y variedad de tratamiento. Lo que hace problemática la polivalencia de un modelo generalista no es el número de turnos de conversación, sino la indeterminación del universo de información sobre el que puede operar y de las finalidades a las que puede servir. En un asistente que recupera exclusivamente de un corpus cerrado y actúa bajo un prompt de sistema que acota su cometido, los turnos sucesivos modifican la formulación de la consulta, pero no el conjunto de datos al que el sistema accede ni la categoría de tratamiento que realiza: la finalidad la fija el corpus, no la interacción.

Subsiste, con todo, un margen residual que conviene reconocer. Nada impide materialmente que un usuario formule una consulta ajena al dominio asignado y que el modelo responda apoyándose en su conocimiento paramétrico en lugar de en el corpus recuperado. Ese supuesto plantea un problema de exactitud: el sistema genera texto sin respaldo documental. No plantea, en cambio, un problema de finalidad, porque el sistema sigue sin tratar datos personales de la entidad para un fin distinto del que tiene asignado. Es el riesgo que mitigan las medidas de configuración descritas más adelante. La frontera entre la herramienta acotada y el asistente versátil no es, por tanto, absoluta. Lo que la arquitectura garantiza es que cruzarla no comprometa la localización del dato ni el perímetro del corpus, de los que dependen las garantías examinadas a continuación.

5.2 Cumplimiento desde el diseño, artículo por artículo

El interés del caso para el presente trabajo reside en que las garantías que el modelo contractual confía a la voluntad de las partes aparecen aquí incorporadas como propiedades de la arquitectura. El contraste con los preceptos del RGPD permite mostrarlo con precisión.

En relación con los principios del artículo 5, la ejecución local satisface el principio de integridad y confidencialidad del apartado 1.f) por una vía distinta de la contractual: el dato no abandona el perímetro físico y lógico de la administración, de modo que la confidencialidad no depende de que un tercero respete lo pactado, sino de que el procesamiento no se externaliza en absoluto.

El principio de limitación de la finalidad se refuerza mediante la especialización: una aplicación de tarea única no es un sistema polivalente al que haya que imponer restricciones de uso desde fuera, sino un sistema cuya finalidad está acotada por diseño, lo que aproxima su comportamiento al de un modelo especializado y reduce la indeterminación funcional que se identificó como cuarta tensión del modelo contractual.

Este desplazamiento de la finalidad desde la imposición externa hacia el diseño se aprecia con particular claridad en el asistente de contratación pública. La herramienta opera mediante recuperación aumentada sobre un corpus documental cerrado, depurado y fiscalizado jurídicamente por la propia institución, integrado de forma exclusiva por la Ley 9/2017 de Contratos del Sector Público, los pliegos tipo aprobados por la entidad y la doctrina de los tribunales administrativos de recursos contractuales. El sistema no crea derecho ni emite opiniones: recupera, sintetiza y referencia la norma vigente citando en cada respuesta la fuente exacta —el artículo, el inciso o la resolución— de la que extrae la información. Esta acotación del corpus cumple una doble función jurídica. Por un lado, materializa el principio de limitación de la finalidad, porque el asistente no puede operar fuera del dominio normativo que tiene asignado. Por otro, la cita obligada de la fuente reduce el riesgo de generación de contenido inexacto, lo que conecta con el principio de exactitud del artículo 5.1.d): la respuesta es verificable contra la norma de la que dice proceder, de modo que el empleado público puede contrastarla en lugar de confiar en ella ciegamente.

El riesgo de que el modelo construya una cita inexistente no se elimina por completo, pues ningún sistema basado en modelos de lenguaje puede garantizarlo de forma absoluta, pero sí se mitiga mediante una combinación de medidas de configuración sobre el motor de recuperación. El empleo de prompts de sistema restrictivos, que instruyen al modelo para responder únicamente a partir del contexto recuperado, junto con el ajuste de parámetros de inferencia como una temperatura reducida y la limitación del muestreo de candidatos, restringe el margen del modelo para completar

la respuesta con conocimiento latente ajeno al corpus. La fidelidad estricta a la fuente recuperada deja así de confiarse a la fortuna del modelo y pasa a depender de una configuración deliberada del sistema, coherente con la lógica de garantía incorporada al diseño que vertebra este análisis.

La consecuencia para el control interno merece subrayarse, porque enlaza con la línea argumental ya reconocida en sede de control público. Al fundamentar cada salida en una referencia normativa precisa y al mantener la trazabilidad documental del proceso, el asistente facilita la posterior labor de fiscalización de la función interventora y desplaza el control desde el terreno reactivo hacia el preventivo. La auditabilidad que el código abierto hace posible en el plano de la seguridad se proyecta así sobre el plano de la legalidad material: el sistema no solo puede inspeccionarse, sino que documenta el fundamento normativo de sus propias propuestas. Frente al escenario de caja negra, en el que el órgano de control carece de acceso a la ponderación de variables y no puede reproducir el razonamiento de la máquina, la arquitectura analizada preserva la cadena de custodia del procedimiento administrativo.

El tratamiento de categorías especiales de datos del artículo 9 constituye el banco de pruebas más exigente. El gestor de correo electrónico procesa comunicaciones que con frecuencia contienen datos de salud, situación socioeconómica, información fiscal o circunstancias de violencia de género, en áreas como los servicios sociales o la recaudación. Procesar estos correos mediante un servicio comercial en la nube exigiría exportar datos del artículo 9 a infraestructuras de terceros, con la consiguiente activación de evaluaciones de impacto y la asunción de riesgos residuales de difícil mitigación. En el modelo analizado, el procesamiento se realiza in situ: los vectores semánticos de los correos se generan y consultan en la memoria local de los servidores propios, sin que el dato sensible salga del perímetro. La garantía del artículo 9 no se construye sobre una cláusula que prohíbe un uso, sino sobre una arquitectura que impide físicamente la salida del dato.

El cumplimiento del artículo 22 del RGPD y del artículo 14 del RIA se materializa como límite técnico, no como directriz organizativa. El artículo 22 del RGPD garantiza al ciudadano el derecho a obtener intervención humana ante decisiones automatizadas con efectos jurídicos; el artículo 14 del RIA exige que el operador diseñe el sistema con supervisión humana efectiva. El ecosistema MencIA satisface ambas exigencias por la misma vía arquitectónica: los asistentes y aplicaciones operan en entornos de lectura, análisis y propuesta, sin capacidad de actuación sobre los sistemas de registro y tramitación de expedientes. La máquina propone borradores estructurados; la validación final y la asunción de responsabilidad recaen, de forma indelegable, en el empleado público identificado y autenticado. La supervisión humana resulta de la propia arquitectura de las aplicaciones, que carecen de integración de escritura con estos sistemas de registro y tramitación. No disponen de interfaz para incorporar asientos, notificar resoluciones, ordenar disposiciones de fondos ni firmar actos administrativos, de modo que la automatización de decisiones con efectos jurídicos no es algo que el sistema deba abstenerse de hacer, sino algo para lo que carece de capacidad técnica. La diferencia importa jurídicamente: una restricción de permisos puede revertirse mediante una decisión de configuración, mientras que la ausencia de la capacidad misma exigiría un desarrollo deliberado.

Respecto del artículo 25, el caso ilustra con nitidez la diferencia entre protección por diseño y protección por contrato. Las decisiones de arquitectura —ejecución local, código abierto inspeccionable, corpus cerrado mediante recuperación aumentada, ausencia de permisos de ejecución autónoma— constituyen medidas técnicas adoptadas en el momento de determinar los medios del tratamiento, que es exactamente el momento que el precepto identifica. La protección de datos deja de ser un compromiso documental añadido al sistema para convertirse en una propiedad del sistema.

En cuanto al artículo 32, la posibilidad de inspección del código abierto permite que la función interventora y la persona delegada de protección de datos examinen el funcionamiento real del sistema, superando la disociación entre plano documental y plano técnico que se señaló como segunda tensión del modelo contractual. La seguridad deja de auditarse solo sobre la descripción del sistema para poder auditarse sobre el sistema mismo.

Finalmente, en relación con los artículos 44 y siguientes, la consecuencia es la más directa: al no externalizarse el procesamiento de información sensible, no se producen transferencias internacionales de esos datos. El problema de la extraterritorialidad estructural, tercera tensión del modelo contractual, no se mitiga mediante garantías reforzadas, sino que se elimina en su origen porque el supuesto de hecho que lo desencadena —la salida del dato hacia una infraestructura de terceros— no llega a producirse.

5.3 Una prueba de validez prospectiva: la arquitectura ante la inteligencia artificial agéntica

El análisis precedente se refiere a un despliegue que es, en su estado actual, mayoritariamente no agéntico: las herramientas operan con autonomía mínima y bajo supervisión humana directa. Cabe sin embargo someter la arquitectura a una prueba prospectiva, preguntando si conservaría su capacidad de garantía en caso de evolucionar hacia configuraciones de mayor autonomía. Las orientaciones de la autoridad de control sobre inteligencia artificial agéntica ofrecen el criterio para esa prueba.

Dichas orientaciones formulan un umbral de mínimos, inspirado en la denominada regla de los dos, según el cual resulta inaceptable una configuración que combine simultáneamente tres capacidades: procesar automáticamente información no controlada, acceder a información sensible y ejecutar acciones automáticas con efecto dentro o fuera de la organización. El ejemplo que la propia autoridad emplea para ilustrar la configuración prohibida es, precisamente, un agente de respuesta automática a mensajes de correo electrónico.[10]

La relevancia de esta coincidencia es considerable, porque el gestor de correo es también el caso emblemático del artefacto analizado. El contraste muestra que la arquitectura neutraliza dos de los tres vértices de la regla por diseño. El sistema no ejecuta acciones automáticas con efecto jurídico, ya que el principio de administración aumentada reserva toda acción dispositiva al empleado público; y el acceso a información sensible se produce dentro de un perímetro cerrado, sin exposición a infraestructuras externas que pudieran introducir contenido no controlado. La configuración que la autoridad señala como inaceptable se desactiva, así, no mediante un control añadido, sino por la propia disposición arquitectónica del sistema.

Conviene precisar el fundamento de esta neutralización. La supervisión humana obligatoria no es en MencIA una restricción de configuración que una decisión posterior pudiera levantar; es la consecuencia de que el sistema carece de permisos lógicos para encadenar tareas de forma autónoma. No existe en la arquitectura la capa que un agente necesitaría para planificar y ejecutar la secuencia siguiente sin intervención: el modelo propone, pero el inicio de cualquier acción con efecto real requiere una acción positiva del empleado. Eso significa que la cadena característica de la IA agéntica no se interrumpe en MencIA, sino que estructuralmente no puede constituirse. Esta diferencia, entre impedir la cadena y cortarla, es jurídicamente relevante porque la primera garantía es inmune a la presión de uso y a los cambios de configuración; la segunda, no.

Un segundo aspecto que las orientaciones destacan es la repetibilidad del comportamiento, que la autoridad considera no inherente a la tecnología sino dependiente de que exista control sobre las fuentes, las versiones de los modelos y la planificación de tareas. La arquitectura analizada satisface esas condiciones: las versiones de los modelos están fijadas y bajo control de la administración, las fuentes de conocimiento se acotan mediante recuperación aumentada sobre corpus cerrados, y la lógica de cada herramienta de tarea única equivale a una cadena de razonamiento predefinida por el administrador, frente a la planificación autónoma que la propia autoridad señala como fuente de comportamiento no repetible. El paradigma de tarea única puede así reinterpretarse, en clave jurídica, como una contención deliberada de la autonomía que sostiene la trazabilidad y la auditabilidad de las operaciones.

5.4 Las limitaciones de la fase inicial

El rigor del análisis exige declarar sus límites. El artefacto se encuentra en una fase de laboratorio y no ha sido objeto de una implantación escalada al conjunto de la provincia ni de una evaluación de impacto consolidada. El marco formal de gobernanza —código ético, auditoría sistematizada, plan de alfabetización conforme al artículo 4 del RIA— está diseñado[11] pero pendiente de implementación completa, de modo que algunas de las garantías que aquí se describen como propiedades de la arquitectura conviven todavía con procesos de gobernanza en construcción.

Esta circunstancia, lejos de invalidar el análisis, lo precisa. Lo que el caso demuestra no es que un modelo de soberanía arquitectónica resuelva por sí solo todas las exigencias del RGPD, sino algo más acotado y, por ello, más defendible: que determinadas garantías que el modelo contractual confía a la voluntad de las partes pueden incorporarse como propiedades estructurales del sistema, y que esa incorporación es técnicamente factible en una administración con recursos limitados. La verificación

[10] Agencia Española de Protección de Datos, *Inteligencia artificial agéntica*, v1.2, apartado IV.G, "Gestión del riesgo", subapartado "Regla de 2".

[11] La hoja de ruta de gobernanza se recoge en *MencIA: Estrategia de Inteligencia Artificial de la Diputación de Lugo*, aprobada por el Pleno de la Diputación de Lugo el 26 de mayo de 2026.

de que tales garantías se sostienen a escala, y su integración en un marco de gobernanza plenamente desplegado, constituyen el objeto de las fases posteriores de la estrategia y exceden el alcance de esta prueba de concepto.

6. Discusión y limitaciones

6.1 El riesgo de absolutizar la soberanía arquitectónica

El argumento desarrollado podría deslizarse hacia una conclusión que conviene desactivar expresamente: que la internalización de la infraestructura resolvería por sí sola las exigencias de protección de datos. No es así. La ejecución local elimina ciertos riesgos en su origen, pero introduce otros que se desplazan hacia la propia organización. Quien asume la soberanía del dato asume también, de forma indivisible, la responsabilidad íntegra sobre la seguridad de la infraestructura donde ese dato reside. La fortificación de los servidores, la segmentación de las redes, la gestión de identidades y privilegios, la detección de intrusiones y la respuesta ante incidentes dejan de ser obligaciones del proveedor para convertirse en cargas directas de la administración. Una arquitectura soberana mal asegurada no es más garantista que un servicio externo bien contratado; puede serlo menos. La soberanía arquitectónica no es, por tanto, una garantía que opere automáticamente por el hecho de ubicar el procesamiento en casa, sino una garantía condicionada a que la organización posea la capacidad técnica de defender lo que ha internalizado.

6.2 Dependencias residuales

El modelo tampoco logra una autarquía tecnológica completa, ni la pretende. La propia estrategia se autodefine mediante el concepto de soberanía realista, que reconoce sus límites: la administración no fabrica el hardware de procesamiento, no entrena desde cero los modelos fundacionales que ejecuta y puede recurrir a modelos comerciales para tareas no críticas que no involucran datos sensibles. Conviene precisar, además, que la disyuntiva no se reduce a la oposición entre modelos abiertos en infraestructura propia y modelos comerciales en la nube. Existen modelos privativos desplegables en infraestructura local, lo que neutralizaría la dimensión extraterritorial del problema; sin embargo, sus esquemas de licenciamiento restrictivo y su naturaleza de caja negra impiden auditar la lógica del sistema, de modo que reproducen, ya en el plano de la auditabilidad, la misma opacidad que se reprochó al modelo contractual. La elección del código abierto no obedece, por tanto, únicamente a la ubicación del procesamiento, sino a que solo él satisface de forma simultánea las exigencias de localización y posibilidad de inspección. Subsisten, con todo, dependencias respecto de las comunidades de desarrollo del software libre empleado y respecto de la cadena de suministro de componentes. La honestidad analítica exige reconocer que la soberanía arquitectónica es un gradiente, no un absoluto: lo que el modelo asegura no es la independencia total, sino el control efectivo sobre la capa donde reside la información sensible y sobre las reglas conforme a las cuales se procesa. La internalización del procesamiento produce además un efecto colateral ajeno al objeto de este trabajo, pero coherente con él: al suprimir el tránsito de información hacia infraestructuras remotas, elimina también el consumo energético asociado a ese tránsito y al procesamiento en centros de datos de terceros. Pretender más sería incurrir en la retórica de la autosuficiencia que el propio proyecto rechaza.

6.3 Auditabilidad necesaria, no suficiente

La posibilidad de que el código abierto pueda ser objeto de inspección, es condición de la rendición de cuentas, pero no la agota. Que un sistema pueda auditarse no significa que se audite de forma efectiva, ni que la organización disponga de los perfiles capaces de hacerlo. La trazabilidad de las operaciones, la repetibilidad de las inferencias y la documentación de los flujos de datos requieren, para traducirse en garantía real, una función de control dotada de competencia técnica y de los medios para ejercerla, así como la participación cualificada de la persona delegada de protección de datos en las decisiones de diseño. La auditabilidad es una propiedad habilitante: abre la posibilidad del control, pero su ejercicio efectivo depende de una gobernanza que, en el caso analizado, está diseñada pero pendiente de despliegue completo. Esta distinción entre la posibilidad técnica del control y su realización organizativa es la que impide presentar el modelo como una solución cerrada y obliga a situarlo en el terreno, más modesto y más sólido, de la prueba de concepto.

A esta distinción se añade una condición material que el caso analizado todavía no satisface. Para que la posibilidad de inspección se convierta en verificación disponible para terceros, y no solo para los órganos de control de la propia entidad, sería preciso que el código de las aplicaciones en pro-

ducción se publicase bajo titularidad institucional, en versiones restringidas a la ejecución sobre modelos locales y depuradas de cualquier referencia a la topología interna de la organización. Mientras ese paso no se complete, la auditabilidad del sistema es interna y su verificación externa permanece indisponible, lo que confirma el carácter de prueba de concepto que aquí se reivindica.

6.4 Una línea prospectiva: la inteligencia artificial como tecnología de mejora de la privacidad

El análisis sugiere, finalmente, una inversión de la perspectiva habitual. La inteligencia artificial suele presentarse como un riesgo que la protección de datos debe contener; el caso examinado apunta a que, bajo determinadas condiciones arquitectónicas, puede operar también como instrumento de garantía. La autoridad de control admite que un sistema agéntico bien diseñado puede comportarse como una tecnología de mejora de la privacidad, por ejemplo, cuando se emplean modelos de lenguaje pequeños en etapas intermedias de un flujo para categorizar, depurar y minimizar la información antes de su tratamiento ulterior, o para exponer a la intervención humana únicamente datos anonimizados y evitar así sesgos en la decisión. La arquitectura soberana es el presupuesto que hace viables estas funciones, porque solo cabe interponer filtros de minimización en un flujo cuyo recorrido completo se controla. Esta posibilidad excede el estado actual del artefacto, que no ha desarrollado todavía tales mecanismos, pero define una dirección de evolución coherente con el modelo y con las orientaciones de la autoridad, en la que la protección de datos desde el diseño dejaría de ser una restricción impuesta al sistema para convertirse en una función que el propio sistema desempeña.

7. Conclusiones

El décimo aniversario del RGPD ofrece la ocasión para revisar los presupuestos sobre los que se edificó su modelo de garantías. Este trabajo ha sostenido que ese modelo, de naturaleza esencialmente contractual, presupone tecnologías acotadas, finalidades determinables, encargados identificables y sistemas inspeccionables, y que la inteligencia artificial generativa comercial tensiona esos presupuestos en cuatro puntos: la opacidad del entrenamiento, la auditabilidad limitada, la extraterritorialidad estructural y la polivalencia funcional. Esa tensión no invalida el modelo contractual, que conserva su vigencia para numerosos tratamientos, pero revela su fatiga cuando se proyecta sobre el tratamiento de categorías especiales de datos mediante sistemas opacos alojados en infraestructuras de terceros.

Frente a esa fatiga, la noción de soberanía arquitectónica propone un complemento de naturaleza distinta: incorporar a la propia ingeniería del sistema las garantías que el modelo contractual confía a la voluntad de las partes. El examen del ecosistema MencIA ha mostrado que esta incorporación es técnicamente factible en una administración con recursos limitados, y que se traduce en propiedades verificables artículo por artículo. La ejecución local mantiene la información sensible del artículo 9 dentro del perímetro institucional sin necesidad de exportarla; la especialización de las herramientas refuerza la limitación de la finalidad y la cita obligada de fuentes sostiene la exactitud; la ausencia de permisos de ejecución autónoma satisface simultáneamente la obligación de supervisión humana del artículo 14 del RIA y el derecho a la intervención humana del artículo 22 del RGPD, ambos como límite técnico y no como mera directriz; la posibilidad de inspección del código abierto habilita la auditoría real exigida por el artículo 32; y la no externalización elimina en su origen el problema de las transferencias internacionales. Todo ello constituye una traducción operativa del mandato de protección de datos desde el diseño del artículo 25.

La autoridad de control española, al resolver para sus propios procesos administrativos que los casos de uso sensibles se implementen en sistemas internos o ad hoc, ha situado el centro de gravedad del cumplimiento en la arquitectura antes que en la cláusula. Se trata de un instrumento autoorganizativo y no de una recomendación de alcance general, pero su procedencia lo convierte en un indicio cualificado, que sus orientaciones posteriores sobre inteligencia artificial agéntica, estas sí dirigidas al sector, vienen a confirmar. La prueba prospectiva realizada muestra que la arquitectura analizada neutraliza por diseño las configuraciones que la autoridad considera de riesgo inaceptable, lo que sugiere que un modelo concebido para la inteligencia artificial generativa conserva su capacidad de garantía frente a tecnologías de mayor autonomía.

La aportación se ofrece con sus límites declarados. El artefacto examinado se encuentra en una fase inicial de laboratorio y su marco de gobernanza está diseñado, pero no plenamente desplegado, de modo que los resultados tienen valor de prueba de concepto operativa y no de evaluación de impacto consolidada. La soberanía arquitectónica no opera automáticamente por el hecho de internalizar la infraestructura, sino que queda condicionada a la capacidad de la organización para asegurar lo que ha internalizado y para ejercer de forma efectiva el control que la auditabilidad hace posible. Re-

conocer estos límites no debilita la tesis, sino que la sitúa en su terreno propio: el de la demostración de que determinadas garantías de protección de datos pueden dejar de confiarse a la palabra dada para inscribirse en la estructura del sistema, y que esa inscripción está al alcance de una administración local que decida asumir el control de su propia infraestructura.

Bibliografía

Arias Lombardero, José Antonio. "Gobierno local inteligente y soberano: una propuesta estratégica para la implementación de IA de código abierto en la provincia de Lugo". *Revista digital CEMCI*, n.º 67 (2025). https://revista.cemci.org/numero-67/pdf/trabajos-de-evaluacion-1-gobierno-local-inteligente-y-soberano-una-propuesta-estrategica-para-la-implementacion-de-ia-de-codigo-abierto-en-la-provincia-de-lugo.pdf

Cotino Hueso, Lorenzo. "Qué concreta transparencia e información de algoritmos e inteligencia artificial es la debida". *Revista Española de la Transparencia*, n.º 16 (2023): 17-63. DOI 10.51915/ret.272.

Diputación de Lugo. *MencIA: Estrategia de Inteligencia Artificial de la Diputación de Lugo*. Aprobada por el Pleno el 26 de mayo de 2026. https://deputacionlugo.gal/es/a-deputacion/mencia

Rivera Capón, Pablo, José Antonio Arias Lombardero y Nicolás Fariña Verea. "Gobernanza algorítmica local: diseño e implementación de un stack de IA soberana en código abierto para entidades supramunicipales". *IDP. Revista de Internet, Derecho y Política*, n.º 43 (2025). DOI 10.7238/idp.v0i43.9800300.

Veale, Michael e Irina Brass. "Administration by Algorithm? Public Management Meets Public Sector Machine Learning". En *Algorithmic Regulation*, editado por Karen Yeung y Martin Lodge, 121-149. Oxford: Oxford University Press, 2019. DOI 10.1093/oso/9780198838494.003.0006.