Gandia Tech, Aplicaciones y Datos S.L

Checklist de Auditoría de Gobernanza de un Agente IA

Marco de referencia: Reglamento (UE) 2024/1689 (AI Act) + Guía AEPD sobre Calidad del Dato en IA/ML

Introducción

Guía introductoria y manual de aplicación

1. Presentación y marco regulatorio

La integración acelerada de Agentes de Inteligencia Artificial (Sistemas de IA Autónomos / Agentic AI) en los procesos operativos, decisionales y de atención al cliente de las organizaciones plantea retos jurídicos, técnicos y éticos sin precedentes. A diferencia de los modelos de IA estáticos o puramente generativos, un agente de IA interactúa con su entorno, ejecuta herramientas de forma autónoma (como llamadas a API, consultas a bases de datos y ejecución de código) y toma decisiones con diversos niveles de supervisión. Esta capacidad de agencia amplifica la necesidad de establecer un marco riguroso de gobernanza, supervisión humana y control de riesgos.

Este documento de auditoría ha sido diseñado como un instrumento técnico-jurídico práctico para evaluar, certificar y documentar el nivel de cumplimiento normativo y la solidez operativa de un agente de IA. Su diseño se fundamenta de forma directa en dos pilares normativos e institucionales clave dentro del contexto europeo y español:

El Reglamento (UE) 2024/1689 de Inteligencia Artificial (AI Act): Marco europeo que establece obligaciones proporcionales según el nivel de riesgo del sistema de IA, abarcando desde la prohibición de prácticas inaceptables y la regulación estricta de sistemas de alto riesgo, hasta los deberes de transparencia para la IA de propósito general e interactiva (Art. 50).

La Guía sobre Calidad del Dato en Aprendizaje Automático e Inteligencia Artificial de la Agencia Española de Protección de Datos (AEPD): Criterios directores centrados en la gobernanza del dato, la prevención de sesgos, la exactitud, el ciclo de vida del dato (entrenamiento, ajuste, entrada y salida) y el cumplimiento de los principios del Reglamento General de Protección de Datos (RGPD, Reglamento UE 2016/679).

2. Estructura y contenido del documento: ¿qué encontrará el usuario?

El checklist está estructurado en nueve bloques funcionales y correlativos, diseñados para cubrir todo el ciclo de vida del agente, desde la conceptualización legal hasta el despliegue operativo y la respuesta ante emergencias:

Sección 0. Ficha del proceso de análisis: El punto de partida de la auditoría. Permite registrar la identificación exacta del agente (versión y módulo), el alcance de la evaluación, los responsables, la periodicidad y los interlocutores clave de la organización (DPO, técnicos, responsables de negocio).

Sección 1. Clasificación del agente y marco legítimo de uso: Bloque orientado a fijar la calificación jurídica del sistema. Permite descartar prácticas prohibidas (Art. 5 AI Act), clasificar el nivel de riesgo (Anexos I y III AI Act), identificar el rol en la cadena de valor (proveedor, desplegador, etc.) y formalizar la base legitimadora del tratamiento conforme al RGPD.

Gandia Tech, Aplicaciones y Datos S.L

Sección 2. Gestión de la calidad del dato: Enfocado en la arquitectura de datos del agente. Evalúa la representatividad y ausencia de sesgos en el entrenamiento o fine-tuning, la seguridad e integridad del contexto y los prompts (prevención de Prompt Injection directo e indirecto) y la contención de "alucinaciones" o datos inexactos en la salida (outputs).

Sección 3. Gestión de riesgos y evaluación de impacto: Analiza la existencia de un Sistema de Gestión de Riesgos continuo (Art. 9 AI Act). Verifica la ejecución y actualización de las evaluaciones obligatorias: la Evaluación de Impacto en Protección de Datos (EIPD/DPIA) y, si procede, la Evaluación de Impacto en Derechos Fundamentales (FRIA/EIAJ).

Sección 4. Control operativo, supervisión humana y planes de parada: Evalúa la gobernanza de la autonomía del agente. Revisa la matriz de delegación de decisiones, los esquemas de intervención humana (Human-in-the-loop, Human-on-the-loop) y los mecanismos técnicos de parada de emergencia (Kill Switch) y retorno a un estado seguro.

Sección 5. Pruebas, validación y red teaming: Comprueba que el agente haya superado pruebas métricas de precisión y robustez antes de su despliegue. Evalúa si el sistema ha sido sometido a pruebas adversarias de ciberseguridad (Red Teaming) para prevenir fuga de datos o escalada de privilegios.

Sección 6. Trazabilidad, registros y explicabilidad: Garantiza la auditabilidad posterior del sistema. Revisa el registro automático de eventos (logging de trazabilidad del pensamiento o chain of thought) (Art. 12 AI Act) y los canales para que los afectados puedan ejercer su derecho a la intervención humana o impugnación (Art. 22 RGPD).

Sección 7. Documentación y evidencias a recopilar: Una lista de verificación (checklist documental) que resume todas las pruebas objetivas (informes, políticas, capturas, matrices, logs) necesarias para respaldar la auditoría.

Sección 8. Dictamen y Conclusiones de la Auditoría: Cierre formal de la evaluación. Permite seleccionar el resultado final (Cumplimiento Total, Cumplimiento Parcial o No Conformidad), detallar el resumen de hallazgos, fijar un plan de acciones correctoras con plazos estrictos y plasmar las firmas de los auditores y responsables.

Sección 9. Guía de uso: Manual operativo integrado dentro del documento para consultar criterios interpretativos específicos punto por punto.

Gandia Tech, Aplicaciones y Datos S.L

Índice de Contenidos

Introducción
0. Ficha del proceso de análisis
1. Clasificación del agente y marco legítimo de uso
    1.1 Clasificación de riesgo (AI Act)
    1.2 Definición del rol organizativo
    1.3 Base jurídica del tratamiento (RGPD/AEPD)
2. Gestión de la calidad del dato
    2.1 Datos de entrenamiento / fine-tuning
    2.2 Datos en entrada (inputs/prompts) y contexto
    2.3 Gobernanza del dato en la salida (outputs)
3. Gestión de riesgos y evaluación de impacto
    3.1 Sistema de gestión de riesgos continuo (Art. 9 AI Act)
    3.2 Evaluación de Impacto en Derechos Fundamentales (FRIA/EIAJ)
    3.3 Evaluación de Impacto en Protección de Datos (EIPD/DPIA, Art. 35 RGPD)
4. Control operativo, supervisión humana y planes de parada
    4.1 Supervisión humana efectiva (Art. 14 AI Act)
    4.2 Planes de parada y parada de emergencia (Kill Switch)
5. Pruebas, validación y red teaming
    5.1 Fase de pruebas y validación (Art. 15 AI Act)
    5.2 Pruebas adversarias (Red Teaming)
6. Trazabilidad, registros y explicabilidad
    6.1 Registro automático de eventos / logging (Art. 12 AI Act)
    6.2 Explicabilidad y transparencia (AEPD / Art. 22 RGPD)
7. Documentación y evidencias a recopilar durante la auditoría
8. Dictamen de la auditoría
9. Guía de uso

Gandia Tech, Aplicaciones y Datos S.L

0. Ficha del proceso de análisis

Campo / Parámetro	Detalle / Datos del Proceso
Nombre y versión del agente IA auditado	
Fecha de la auditoría y periodicidad prevista (Revisión continua o puntual)	
Responsable/equipo que ejecuta la auditoría (Interno / Auditor externo)	
Ámbito de la auditoría (Todo el sistema, un módulo, una nueva funcionalidad, un cambio sustancial)	
Versión del documento de gobernanza vigente sobre el que se audita	
Identificación de la entidad y su(s) rol(es) en la cadena de valor (ver punto 1.2)	
Lista de interlocutores clave (DPO, responsable de IA/compliance, equipo técnico, negocio)	

Gandia Tech, Aplicaciones y Datos S.L

1. Clasificación del agente y marco legítimo de uso

1.1 Clasificación de riesgo (AI Act)

☐ Se ha verificado que el agente no incurre en ninguna práctica prohibida del Art. 5 (manipulación subliminal, explotación de vulnerabilidades, social scoring, etc.)

☐ Se ha evaluado si el agente encaja en algún supuesto de alto riesgo (Anexos I y III): infraestructuras críticas, empleo/selección de personal, acceso a servicios esenciales públicos o privados, justicia/procesos democráticos, identificación biométrica

☐ Si es de alto riesgo: existe registro en la base de datos UE correspondiente y se cumplen los requisitos del Título III (sistema de gestión de calidad, documentación técnica, etc.)

☐ Si el agente interactúa directamente con personas físicas: se ha implementado el deber de transparencia del Art. 50 (aviso claro de que se interactúa con un sistema de IA)

☐ Existe evidencia documental (captura, texto legal, disclaimer) del cumplimiento del deber de transparencia

☐ Se ha revisado si el agente genera o manipula contenido (texto, imagen, audio, vídeo) sujeto a obligaciones de etiquetado como IA generativa

1.2 Definición del rol organizativo

☐ Se ha determinado formalmente si la organización actúa como Proveedor, Desplegador, Importador o Distribuidor

☐ Existe un mapa de responsabilidades asociado a cada rol y a cada fase del ciclo de vida

☐ Si hay terceros implicados (proveedores del modelo base, integradores, subcontratistas), están identificados y sus obligaciones contractuales están documentadas

1.3 Base jurídica del tratamiento (RGPD/AEPD)

☐ Está documentada expresamente la base legitimadora del tratamiento de datos personales que procesa el agente (contrato, interés legítimo, consentimiento, obligación legal, etc.)

☐ El test de ponderación de interés legítimo (si aplica) está documentado y actualizado

☐ Se ha verificado la compatibilidad de finalidades si los datos se reutilizan para entrenar/ajustar el agente

☐ Están identificadas las categorías de datos personales tratadas y si incluyen categorías especiales (Art. 9 RGPD)

Gandia Tech, Aplicaciones y Datos S.L