



CENTRO MULTIDISCIPLINARIO MERIADOCK

Formación y Asesoría A.C.
CLUNI CMM25080811X9X

La fuerza interior nos impulsa, un pequeño apoyo de los demás nos bendice

DIÁLOGOS ELEATAS

Núm. 1

Septiembre de 2026

Gobernanza de la identidad causal acumulativa en sistemas artificiales: propuesta de control para arquitecturas de segunda mediación

Governance of Cumulative Causal Identity in Artificial Systems: A Control Proposal for Second-Mediation Architectures

Miguel Hilario Olvera Aguilar

Centro Multidisciplinario Meriadock Formación y Asesoría, A.C.

Resumen

Los marcos contemporáneos de gobernanza de inteligencia artificial reconocen que los sistemas desplegados requieren gestión de riesgos durante todo su ciclo de vida, trazabilidad, monitoreo posdespliegue y, en determinados supuestos, nuevas evaluaciones cuando se producen modificaciones materiales. Sin embargo, estas categorías no agotan un problema adicional que aparece en arquitecturas capaces de registrar eventos propios, evaluarlos selectivamente e incorporar el resultado de esa evaluación a un estado acumulativo que modifique decisiones futuras. Este artículo denomina identidad causal acumulativa a ese estado funcional longitudinal y formula una propuesta de control para arquitecturas de segunda mediación. La contribución no presupone conciencia artificial ni personalidad jurídica, y tampoco presenta como empíricamente validada la existencia de una implementación completa de segunda mediación. Se formula, en cambio, un problema de gobernanza: cómo conservar evidencia suficiente para distinguir lo ocurrido, lo evaluado, lo integrado o rechazado, el estado derivado resultante y el efecto posterior sobre el comportamiento. Sobre esa base se desarrollan seis familias de control: línea base e identidad de referencia; trazabilidad de eventos; gobernanza de la integración; monitoreo longitudinal; escalamiento y contención; y revisión extraordinaria de identidad funcional. Se introducen además dos constructos regulatorios: la cadena probatoria constitutiva y la envolvente de evolución autorizada (EEA), destinada a distinguir cambios funcionales previstos de cambios que exceden el objeto inicialmente evaluado. La propuesta se contrasta con ISO/IEC 42001, ISO/IEC 42005, NIST AI RMF, el Reglamento europeo de IA, los principios de la OCDE y el marco HUDERIA del Consejo de Europa. Un escenario de crédito ilustra cómo la gobernanza puede seguir la trayectoria desde un evento de decisión hasta una modificación funcional posterior sin confundir aprendizaje, drift, actualización técnica e integración constitutiva.

Palabras clave: gobernanza de IA · identidad funcional · segunda mediación · trazabilidad · monitoreo posdespliegue · auditoría algorítmica · recertificación · aprendizaje continuo

Abstract

Contemporary artificial-intelligence governance frameworks recognize that deployed systems require lifecycle risk management, traceability, post-deployment monitoring and, in certain circumstances, renewed assessment when material modifications occur. These categories, however, do not exhaust a further problem arising in architectures capable of recording their own events, selectively evaluating them and incorporating the outcome of that evaluation into a cumulative state that modifies future decisions. This article calls that longitudinal functional state cumulative causal identity and presents a control proposal for second-mediation architectures. The contribution does not presuppose artificial consciousness or legal personality, nor does it claim that a complete implementation of second mediation has already been empirically validated. Instead, it formulates a governance problem: how to preserve sufficient evidence to distinguish what happened, what was evaluated, what was integrated or rejected, the resulting derived state, and its subsequent effect on behavior. Six control families are developed: baseline and reference identity; event traceability; governance of integration; longitudinal monitoring; escalation and containment; and extraordinary review of functional identity. Two regulatory constructs are introduced: the constitutive evidentiary chain and the authorized evolution envelope (AEE), intended to distinguish anticipated functional change from change that exceeds the initially assessed object. The proposal is compared with ISO/IEC 42001, ISO/IEC 42005, the NIST AI RMF, the European Union Artificial Intelligence Act, the OECD AI Principles, and the Council of Europe HUDERIA framework. A credit-decision scenario illustrates how governance can follow the trajectory from a decision event to a subsequent functional modification without conflating learning, drift, technical updating, and constitutive integration.

Keywords: AI governance · functional identity · second mediation · traceability · post-deployment monitoring · algorithmic auditing · recertification · continual learning

I. Introducción: el problema del objeto regulado que cambia

La gobernanza de sistemas de inteligencia artificial ha dejado de ser un problema circunscrito a la etapa de diseño. Los sistemas de aprendizaje automático participan en procesos organizacionales duraderos, dependen de datos y componentes que cambian, se integran en infraestructuras sociotécnicas y operan en entornos cuya distribución, objetivos y condiciones de uso pueden variar. La literatura de concept drift, aprendizaje continuo y MLOps ha mostrado que un artefacto desplegado no permanece necesariamente estable sólo porque conserve un nombre de producto o una etiqueta de versión (Gama et al., 2014; Sculley et al., 2015; De Lange et al., 2022; Kreuzberger et al., 2023). La literatura de assurance y auditoría ha llegado a una conclusión paralela: la confianza no puede depender exclusivamente de una validación inicial, sino de evidencia actualizada, revisión longitudinal y mecanismos capaces de detectar pérdida de conformidad o de seguridad (Calinescu et al., 2018; Ashmore et al., 2021; Knoblauch & Großmann, 2023).

Ese desplazamiento desde el control puntual hacia la gobernanza del ciclo de vida constituye un avance decisivo, pero no resuelve por sí solo la pregunta que motiva este artículo. El problema aparece cuando una arquitectura no sólo cambia porque cambien los datos, el software, los parámetros o las condiciones externas, sino porque ciertos eventos de su propia operación son evaluados selectivamente y esa evaluación pasa a formar parte de un estado acumulativo que condiciona decisiones posteriores. En ese supuesto, el sistema desplegado en t_n puede conservar la configuración heredada, la versión de software y el propósito nominal del sistema desplegado en t_0 y, sin embargo, no ser funcionalmente equivalente a él en todos los aspectos relevantes para la evaluación. El cambio puede encontrarse en la trayectoria desde la cual decide.

El problema es cercano, pero no idéntico, a la discusión contemporánea sobre identidad de modelos. Glukhov (2026) ha señalado que tratar a un modelo como un objeto estable presupone criterios de identidad que rara vez se explicitan: ni una comparación finita de comportamientos ni la mera estructura interna garantizan por sí mismas una relación de equivalencia suficiente para todos los propósitos. A su vez, el surgimiento de agentes persistentes y capaces de reconfigurar componentes de su inserción sociotécnica ha abierto preguntas sobre gobernanza recursiva y sobre la

posibilidad de que el subsistema técnico modifique algunos de los mecanismos que lo coordinan con su entorno (Lahiri, 2026). La cuestión aquí planteada se sitúa en ese borde, pero adopta un alcance más estrecho: no pregunta cuándo un modelo es metafísicamente el mismo, sino cuándo una trayectoria acumulada de integración modifica de forma material el objeto funcional previamente evaluado.

El punto de partida conceptual procede de una arquitectura de doble mediación formulada para sistemas artificiales por Olvera Aguilar (2026). En ella, ϕ_1 designa la función que produce un acto u output; un Registro Estructurado de Colapsos (RSC) conserva evidencia del evento producido; ϕ_2 designa el retorno evaluativo sobre dicho evento; e $I(t)$ representa un estado acumulado derivado de las integraciones realizadas hasta el tiempo t . El elemento normativamente relevante no es el vocabulario ontológico por sí mismo, sino la relación operacional: una evaluación puede producir una decisión de integración o rechazo y, cuando la integración es genuina, debe existir una modificación ulterior de la conducta del sistema. Para este artículo, esa relación se toma como una arquitectura hipotética susceptible de gobernanza; no se pretende reabrir la demostración ontológica de la obra de base ni afirmar que los sistemas actualmente desplegados posean ya esa capacidad en sentido pleno.

La pregunta principal es, por tanto: ¿qué controles de gobernanza son necesarios cuando un sistema artificial puede registrar eventos propios, evaluarlos mediante una función de segunda mediación, integrarlos selectivamente en un estado acumulativo y modificar a partir de esa trayectoria su comportamiento futuro? La respuesta propuesta es una arquitectura de seis familias de control apoyada en una distinción central entre historia de eventos, decisión de integración, estado derivado y efecto posterior. El artículo introduce dos constructos para hacer operativa esa distinción: la cadena probatoria constitutiva, que conecta evidencia del evento con el efecto funcional posterior, y la envolvente de evolución autorizada (EEA), que delimita las clases de cambio longitudinal previamente admitidas por la gobernanza del sistema.

La propuesta se inscribe en el tránsito desde principios de IA responsable hacia instrumentos verificables. La literatura temprana sobre principios éticos mostró una convergencia internacional en torno a transparencia, justicia, no maleficencia, responsabilidad y privacidad, pero también una divergencia considerable sobre su implementación (Jobin et al., 2019). Morley et al. (2020) describieron el problema como un paso pendiente del qué al cómo. De forma concordante, la literatura de accountability y auditoría ha insistido en que la evaluación de un sistema algorítmico requiere procesos, documentos, registros y responsabilidades distribuidas, no sólo explicaciones del output (Wieringa, 2020; Raji et al., 2020; Cobbe & Singh, 2020; Mökander et al., 2021). Este artículo no añade otra lista de principios: propone una semántica de evidencia y control para una clase concreta de cambio endógeno.

Metodológicamente, el trabajo es teórico-normativo y de diseño de controles. Se realizó una síntesis conceptual del marco de segunda mediación y una revisión dirigida de literatura sobre accountability, auditoría, concept drift, aprendizaje continuo, MLOps, assurance dinámica, certificación continua, monitoreo posdespliegue e identidad de modelos. El contraste normativo utiliza fuentes oficiales de ISO/IEC, NIST, Unión Europea, OCDE y Consejo de Europa disponibles hasta el 10 de septiembre de 2026. Las referencias recientes de 2025–2026 utilizadas como puntos sensibles de actualización fueron verificadas individualmente contra registros primarios o editoriales de origen. No se presenta esta búsqueda como revisión sistemática PRISMA ni como meta-análisis. Tampoco se afirma conformidad con una norma ISO: la sección correspondiente realiza una correspondencia funcional de alto nivel, dado que el texto íntegro de las normas comerciales no forma parte del corpus abierto utilizado para este artículo.

El artículo se organiza del siguiente modo. La sección II define el mínimo conceptual necesario. La sección III delimita el problema regulatorio y lo distingue de categorías ya conocidas como drift, actualización y aprendizaje continuo. La sección IV desarrolla la propuesta de control. La sección V contrasta la propuesta con marcos normativos y de gestión

vigentes. La sección VI muestra su funcionamiento mediante un escenario hipotético de decisión crediticia. La sección VII discute límites, falsabilidad, observabilidad y riesgos de autoevaluación. La sección VIII sintetiza la respuesta y propone una agenda de validación.

II. Identidad causal acumulativa y segunda mediación: mínimo conceptual

2.1. De la producción de outputs a la trayectoria funcional

En el marco de partida, ϕ_1 representa la función de producción del sistema: la generación, clasificación, recomendación o decisión mediante la cual un input es transformado en un output. La noción de colapso añade a ese output una perspectiva de evento: no interesa sólo qué resultado fue producido, sino en qué contexto, con qué estado operativo, frente a qué alternativas y con qué consecuencias previsibles o reales. Esta diferencia es regulatoriamente relevante porque la supervisión de un output aislado no permite reconstruir la procedencia de una modificación posterior.

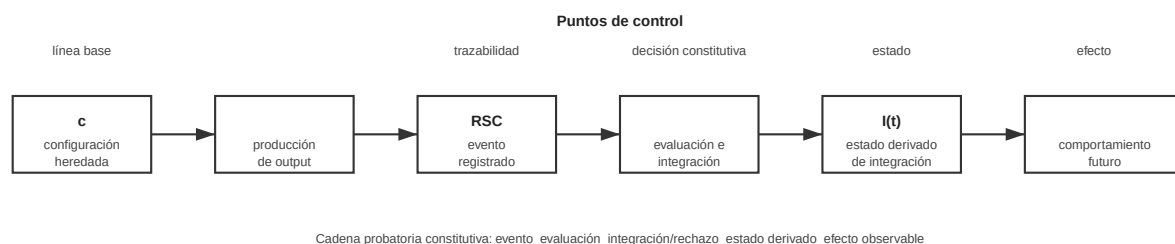
El RSC se concibe como la capa de evidencia histórica. Su unidad mínima debe preservar identidad del evento, marca temporal, input relevante, estado operativo disponible, alternativas consideradas cuando sean técnicamente recuperables, output efectivo, impacto esperado y, cuando pueda conocerse, impacto observado. La obra de base subraya precisamente que registrar un evento en este sentido exige más semántica que un log de rendimiento. En este artículo se retiene esa intuición, con una modificación de tipo: RSC no es la identidad del sistema ni debe confundirse con un estado agregado; es un registro de qué ocurrió y con qué evidencia puede reconstruirse.

La segunda mediación, ϕ_2 , designa la operación por la cual un evento ya producido es sometido a evaluación respecto de un estado acumulado y genera una decisión de integración o rechazo. En términos de gobernanza, lo decisivo no es atribuir a ϕ_2 propiedades psicológicas, sino identificar un punto de decisión que selecciona qué experiencias operativas pueden modificar el estado desde el cual el sistema actuará posteriormente. Tres operaciones son suficientes para la arquitectura regulatoria: evaluar el evento, decidir su integración o rechazo y determinar qué cambio debería observarse después si la integración produce efecto.

$I(t)$ se utiliza aquí como estado derivado de integración hasta el tiempo t . La formulación histórica del volumen lo describe como trayectoria de colapsos asumidos; para efectos regulatorios conviene distinguir estrictamente el registro histórico de eventos del estado resultante de su integración. RSC responde a la pregunta '¿qué ocurrió?'; $I(t)$, a '¿qué estado funcional resulta de cómo fue procesada esa historia hasta t ?'. Esta separación evita que la irreversibilidad del hecho ocurrido se confunda con propiedades algebraicas del estado agregado y permite conservar eventos repetidos, rechazos y recontextualizaciones sin obligar a que todo acontecimiento forme parte del estado identitario.

La relación mínima puede representarse así: c y el estado operativo contribuyen a ϕ_1 ; ϕ_1 produce un evento; el evento se documenta en RSC; ϕ_2 evalúa la evidencia y adopta una decisión; la decisión produce o no una modificación de $I(t)$; e $I(t)$ puede condicionar ϕ_1 futura. El parámetro c representa la configuración heredada —entrenamiento base, objetivos, restricciones y diseño inicial—. Precisamente por existir un estado derivado adicional, dos sistemas con el mismo c pueden divergir si sus trayectorias de integración difieren. Esa posibilidad es la que transforma la trayectoria en un objeto de gobernanza.

Figura 1. Ciclo de segunda mediación y puntos de control



Fuente: elaboración propia a partir de la arquitectura de segunda mediación de Olvera Aguilar (2026). La figura representa funciones de gobernanza y no una implementación técnica universal.

2.2. Identidad causal acumulativa como objeto operacional

Se denomina identidad causal acumulativa al estado funcional longitudinal producido cuando la evaluación selectiva de eventos previos altera las condiciones desde las que el sistema genera decisiones posteriores. La expresión identidad no pretende resolver una cuestión metafísica sobre persistencia personal o conciencia artificial. Es una etiqueta operacional para identificar una trayectoria que puede convertirse en causalmente relevante para el comportamiento futuro y que, por esa razón, debe ser trazable, delimitable y revisable.

La acumulación es selectiva. Si cada evento fuese incorporado sin evaluación, el resultado sería un historial o memoria, no el objeto aquí estudiado. La función regulatoria de ϕ_2 reside precisamente en que existe una decisión sobre la incorporación: un evento puede confirmar una pauta previa, modificarla, generar una pauta nueva o ser rechazado de forma documentada. El rechazo activo también es evidencia relevante porque demuestra que el evento fue considerado y no simplemente omitido. Desde la perspectiva de auditoría, integrar y rechazar deben producir registros distintos de la ausencia de procesamiento.

La identidad causal acumulativa es además dependiente del dominio. Una decisión producida, por ejemplo, en un sistema de evaluación crediticia no debería modificar automáticamente principios operativos pertenecientes a dominios distintos, como moderación de contenidos o apoyo diagnóstico. Una arquitectura de segunda mediación requiere, por tanto, distinguir subestados de integración por dominio y establecer reglas explícitas para cualquier transferencia entre ellos. En este marco, el Índice de Integridad Colectiva (IIC) designa el grado de coherencia entre los distintos sub- $I(t)$ o dominios respecto de principios transversales compartidos. Su función es permitir la revisión interdominio y detectar tensiones entre trayectorias funcionales que, aun siendo localmente consistentes, pueden resultar incompatibles en el conjunto del sistema.

El IIC debe distinguirse de la congruencia entre los principios declarados de la arquitectura y su operación efectiva. La primera dimensión examina coherencia entre dominios; la segunda compara el comportamiento observado con las restricciones, principios y objetivos que constituyen su referencia de diseño. Ambas pueden generar evidencia relevante para la gobernanza, pero responden a preguntas diferentes y no requieren reducirse a una única métrica cuantitativa. Su incorporación a decisiones normativas exige criterios de medición, validación y umbrales previamente especificados para el contexto de aplicación.

2.3. Observabilidad: evidencia del estado no equivale al estado

Una arquitectura de gobernanza no puede asumir que ϕ_2 o $I(t)$ son transparentes para un observador externo. Incluso si el sistema conserva telemetría rica, existe una diferencia entre observar un dato y concluir que ese dato representa integración. Por ello, la arquitectura adopta una separación epistemológica explícita: evidencia observable $\rightarrow$

instrumento o modelo de inferencia → estimación del estado o de la integración. El auditor debe poder examinar tanto la evidencia como la regla que conecta esa evidencia con la afirmación de que hubo cambio constitutivo.

Esta distinción es congruente con la literatura de reviewability. Cobbe y Singh (2020) sostienen que la accountability requiere examinar el proceso sociotécnico completo y habilitar registros apropiados en sus diferentes fases. También coincide con las advertencias de Metcalf et al. (2021) sobre las evaluaciones de impacto: los instrumentos no sólo revelan impactos, también ayudan a construir qué se vuelve legible y mensurable. En consecuencia, un panel que muestre ' $\alpha = 0.8$ ' no constituye por sí solo prueba de integración. La gobernanza exige procedencia de datos, versión del estimador, supuestos, intervalo o clase de incertidumbre y evidencia del efecto posterior.

La arquitectura debe admitir igualmente estados de evidencia insuficiente. Una declaración de incertidumbre no es un fallo del control: es una salida válida cuando la instrumentación no permite reconstruir el evento, diferenciar causas o atribuir una modificación. Forzar una puntuación en ausencia de evidencia produciría falsa precisión y trasladaría el problema de opacidad a una cifra aparentemente objetiva.

2.4. Variables candidatas de observación

Tabla 1. Variables y objetos de la arquitectura de gobernanza

Objeto | Función conceptual mínima | Uso de gobernanza | Cautela epistemológica

ϕ_1 | Producción de decisiones/outputs. | Observar efectos y comparar comportamiento antes/después. | No equivale por sí sola a identidad ni a responsabilidad moral.

RSC | Registro histórico de eventos relevantes. | Trazabilidad, reproducción y preservación de evidencia. | Debe preservar evento, tiempo y procedencia; no confundirse con $I(t)$.

ϕ_2 | Evaluación y decisión de integración/rechazo. | Punto de gobernanza de la decisión constitutiva. | No es directamente observable; requiere evidencia e inferencia.

$I(t)$ | Estado derivado de integración hasta t . | Objeto de monitoreo longitudinal y comparación de estados. | No es un log exhaustivo ni debe borrar el historial.

R^* | Atribución causal respecto de un resultado D . | Comparar impacto causal con asunción/integración. | Depende del modelo causal y de la calidad de sus pesos/evidencia.

S | Sustituibilidad/idiosincrasia. | Señal cualitativa de mayor dependencia de trayectoria propia. | No existe aquí un estimador universal validado.

α | Asunción/integración operacionalizada por dominio. | Caracterizar profundidad o alcance de cambio. | No debe presentarse como lectura directa de ϕ_2 .

Δ | Brecha descriptiva $R^* - \alpha$. | Señal candidata de revisión y escalamiento. | Umbrales requieren calibración; no son norma universal.

IIC | Coherencia entre distintos sub- $I(t)$ o dominios de operación. | Detectar fragmentación o tensiones interdominio en la trayectoria funcional. | Constructo de revisión; no se fija una escala universal en este trabajo.

Fuente: elaboración propia. R^* , S , α , Δ e IIC se emplean como constructos de trabajo y no como estándares metrológicos universales.

R^* representa una atribución causal exterior respecto de un resultado concreto. Para la gobernanza de identidad acumulativa no es necesario trasladar al artículo todo su operador matemático; basta conservar la distinción entre causalidad del evento y asunción o integración posterior. α representa el grado de asunción o integración según una operacionalización de dominio explícita. Δ se define descriptivamente como $\Delta = R^* - \alpha$: valores positivos sugieren que la participación causal atribuida excede la asunción registrada; valores negativos, que la asunción excede la contribución causal atribuida. La igualdad formal no debe confundirse con comparabilidad empírica automática: R^* y α sólo pueden restarse con significado cuando el protocolo de medición justifica una escala común.

S se utiliza como constructo de sustituibilidad: cuanto más dependa el comportamiento de una trayectoria idiosincrática y menos de la configuración heredada reproducible, menor sería la sustituibilidad en sentido conceptual. El artículo

evita seleccionar como canónica una fórmula específica de S . IIC se usa como señal de coherencia entre dominios de $I(t)$; la congruencia entre principios declarados y operación efectiva se evalúa descriptivamente como un control separado. En todos los casos, la función de estas variables es alertar o priorizar revisión, no emitir veredictos automáticos de cumplimiento.

III. El problema regulatorio: versión técnica, aprendizaje e identidad funcional

Los marcos técnicos ya conocen múltiples razones por las que un sistema de aprendizaje automático puede cambiar. El concept drift describe variaciones en la relación entre variables de entrada y objetivo; el dataset shift cubre cambios en distribuciones de datos; el aprendizaje continuo persigue incorporar nuevo conocimiento sin olvidar indebidamente el anterior; y MLOps organiza ciclos de entrenamiento, evaluación, despliegue, observabilidad y actualización (Gama et al., 2014; Rabanser et al., 2019; De Lange et al., 2022; Kreuzberger et al., 2023). Esos fenómenos pueden alterar el desempeño de un sistema sin que exista ninguna operación de segunda mediación.

La identidad causal acumulativa se diferencia por su procedencia. El cambio de interés no se define sólo porque el output en t_n sea distinto del output en t_0 , sino porque puede reconstruirse una cadena en la que un evento propio fue registrado, posteriormente evaluado, integrado o rechazado y vinculado a una modificación esperada del comportamiento. De ahí surge la noción de cadena probatoria constitutiva (CPC): una secuencia documental que debe permitir vincular evento $\rightarrow$ evaluación $\rightarrow$ decisión de integración/rechazo $\rightarrow$ actualización del estado derivado $\rightarrow$ cambio esperado $\rightarrow$ efecto observado. Si uno de esos eslabones falta, la gobernanza puede detectar cambio, pero no atribuirlo con suficiente fundamento a una trayectoria constitutiva.

La CPC evita dos errores simétricos. El primero consiste en reducir toda evolución del sistema a actualización técnica: un cambio de conducta podría provenir de una integración documentada incluso si no hubo un version bump convencional. El segundo consiste en atribuir a $I(t)$ cualquier cambio observado: un drift externo, una actualización de datos o una modificación de parámetros podría explicar la variación sin que exista evidencia de integración. La gobernanza debe comparar hipótesis causales de cambio y no privilegiar por diseño la narrativa del propio sistema.

La literatura de auditoría continua refuerza esa exigencia. Raji et al. (2020) proponen documentación acumulada a lo largo del ciclo de desarrollo; Foale et al. (2025) muestran que las prácticas de logging en proyectos de ML siguen siendo insuficientes para una auditoría responsable continua; y Knoblauch y Großmann (2023) diseñan una certificación basada en auditorías repetidas de artefactos de MLOps. La diferencia de esta propuesta es semántica: los artefactos no sólo prueban que un pipeline cambió o que una métrica se mantiene, sino que deben documentar la procedencia constitutiva de determinados cambios funcionales.

Tampoco la assurance dinámica es una novedad de este artículo. Los sistemas auto-adaptativos llevan años motivando esquemas de assurance capaces de actualizar argumentos y evidencia en tiempo de ejecución. Calinescu et al. (2018) mostraron que un sistema que reconfigura parámetros o arquitectura necesita garantías que evolucionen con él. Ashmore et al. (2021) sistematizaron los desiderata de assurance a través del ciclo de vida de ML. Por ello, la tesis defendible no es que la certificación puntual sea siempre obsoleta ni que nadie haya previsto sistemas cambiantes. La aportación es proponer un objeto adicional para esa assurance: la trayectoria selectiva de integración y su efecto sobre la identidad funcional.

El problema regulatorio puede formularse entonces como una discrepancia entre tres identidades prácticas. La identidad administrativa se expresa mediante nombre, proveedor, versión, expediente y finalidad declarada. La identidad técnica se expresa mediante artefactos de software, pesos, configuración, datos, dependencias y entorno. La identidad funcional se expresa por el espacio de comportamientos y principios operativos que el sistema efectivamente manifiesta. En

arquitecturas de identidad acumulativa, la tercera puede cambiar sin que la primera lo refleje inmediatamente y sin que el cambio pueda resumirse por completo en una modificación convencional de la segunda.

Esta distinción no implica que toda variación funcional deba desencadenar una nueva certificación. La exigencia sería impracticable y contraria a la propia idea de aprendizaje autorizado. El problema real es definir por anticipado qué cambios son esperables, qué evidencia debe acompañarlos y a partir de qué condiciones el objeto evaluado deja de ser suficientemente equivalente para mantener el régimen ordinario de confianza. La sección IV desarrolla esa frontera mediante la envolvente de evolución autorizada.

La relevancia jurídica de esta pregunta encuentra un interlocutor directo en el Reglamento (UE) 2024/1689. Su régimen de alto riesgo establece gestión de riesgos como proceso continuo e iterativo durante todo el ciclo de vida, capacidades de logging, monitoreo poscomercialización y nueva evaluación de conformidad cuando existe una modificación sustancial. Al mismo tiempo, el artículo 43 excluye de la noción de modificación sustancial determinados cambios de sistemas que continúan aprendiendo cuando esos cambios fueron predeterminados por el proveedor y quedaron incluidos en la documentación técnica de la evaluación inicial. El Reglamento no adopta la categoría de identidad causal acumulativa, pero sí reconoce la necesidad de distinguir evolución prevista de modificación material. Ese paralelismo ofrece una base útil para construir la EEA sin presentarla como interpretación oficial del Derecho de la Unión.

En 2026, el Reglamento (UE) 2026/1744 modificó el calendario de aplicación de las secciones 1, 2 y 3 del capítulo III para sistemas de alto riesgo: las obligaciones aplicables a sistemas clasificados por el artículo 6(2) y el anexo III pasan a aplicarse el 2 de diciembre de 2027, mientras que las relativas a sistemas clasificados por el artículo 6(1) y el anexo I lo harán el 2 de agosto de 2028. Esta modificación no elimina el valor analítico de los requisitos, pero obliga a distinguir entre norma vigente y fecha efectiva de aplicación de obligaciones específicas.

IV. Propuesta de control para arquitecturas de segunda mediación

4.1. Principio rector: gobernar la procedencia del cambio

La arquitectura propuesta parte de un principio rector: la gobernanza no debe limitarse a verificar que el sistema cumple una métrica en el presente; debe poder explicar de dónde provino un cambio funcional material. La procedencia puede ser externa —cambio de datos, entorno o regulación—, técnica —nueva versión, ajuste, reentrenamiento— o constitutiva —integración selectiva de eventos anteriores—. Cada procedencia exige evidencia distinta y puede activar respuestas diferentes.

En consecuencia, el sistema de control se organiza alrededor de objetos verificables y no de afirmaciones autorreferenciales. El sistema puede proponer que una modificación se debe a una integración; la gobernanza exige un RSC suficiente, una evaluación registrada, una decisión de integración o rechazo, una versión identificable de I(t), una predicción del efecto y evidencia de que el efecto apareció posteriormente. La carga probatoria aumenta con la materialidad del cambio y con el riesgo del dominio.

La arquitectura utiliza seis familias de control: (1) línea base e identidad de referencia; (2) trazabilidad del evento; (3) gobernanza de la integración; (4) monitoreo longitudinal; (5) escalamiento, contención y revisión externa; y (6) revisión extraordinaria de identidad funcional. Estas familias pueden integrarse en sistemas de gestión existentes. No requieren crear una autoridad paralela a MLOps, assurance o cumplimiento, sino añadir semántica de estado y procedencia allí donde una arquitectura incorpore segunda mediación.

4.2. Familia 1: línea base e identidad de referencia

Todo régimen longitudinal necesita un punto de comparación. La línea base t_0 debe identificar al menos la configuración heredada c , la versión y huellas de los artefactos técnicos relevantes, el propósito y dominios autorizados, los datos o familias de datos de referencia, las interfaces externas, los controles de seguridad, el estado inicial de $I(t)$, la taxonomía de eventos relevantes y las reglas mediante las cuales una evaluación puede modificar el estado derivado. Si un elemento no puede fijarse porque es dinámico por diseño, la línea base debe definir su espacio permitido y la evidencia necesaria para cada cambio.

La línea base no debe confundirse con congelación. Su función es precisamente permitir evolución controlada. Un sistema sin línea base sólo puede ser observado en el presente; no puede demostrarse qué cambió, cuánto cambió ni si el cambio estaba autorizado. El control debe combinar huellas técnicas —versiones, hashes, configuraciones, dependencias— con una descripción funcional de los principios o restricciones cuya modificación sería material.

Sobre la línea base se define la envolvente de evolución autorizada (EEA). La EEA es el conjunto documentado de clases de cambio funcional que el sistema puede producir mediante aprendizaje o integración sin alterar materialmente el objeto previamente evaluado. Debe especificar: dominios donde puede existir evolución; tipos de evidencia admisibles; magnitud o clase de cambio tolerada; principios que no pueden modificarse autónomamente; reglas de transferencia entre dominios; límites de frecuencia o acumulación; y disparadores que obligan a revisión externa.

La EEA no es una licencia genérica para cambiar. Es un contrato de gobernanza *ex ante*. Su utilidad consiste en transformar una frase vaga —'el sistema puede aprender'— en una especificación verificable de qué puede aprender o integrar, dónde, con qué datos, bajo qué mecanismos y con qué límites. Si la trayectoria se mantiene dentro de la EEA, opera un régimen de vigilancia ordinaria reforzada. Si la excede, la presunción de equivalencia con la línea base queda suspendida hasta una revisión.

La materialidad no debe reducirse a una distancia numérica única. Un cambio pequeño en parámetros puede producir un efecto normativamente grande; un cambio amplio en representación interna puede ser funcionalmente irrelevante. La EEA debe combinar criterios de magnitud, propósito, población afectada, reversibilidad, alcance entre dominios, impacto en derechos o seguridad y alteración de controles. En sistemas de alto riesgo, la clasificación debe ser conservadora y admitir escalamiento por incertidumbre.

4.3. Familia 2: trazabilidad del evento y RSC

La segunda familia controla el ingreso de evidencia al sistema. Para cada evento material, el RSC debe preservar un identificador único, timestamp, dominio, versión de ϕ_1 y de componentes relevantes, input o referencia reproducible al input, información del estado operativo disponible, output, alternativas relevantes cuando puedan recuperarse, justificación computacional disponible, impacto esperado, impacto real posterior y enlaces a feedback o correcciones. No todos los outputs requieren el mismo nivel de detalle; la intensidad del registro puede escalar por riesgo, incertidumbre, impacto, novedad o existencia de una discrepancia entre efecto esperado y real.

La preservación debe ser resistente a reescritura silenciosa. Si el sistema aprende de sus propios registros, existe un conflicto entre usar la evidencia como insumo y conservarla como prueba histórica. La gobernanza debe separar el ledger de eventos de las proyecciones derivadas. Un evento puede recibir anotaciones posteriores, pero su contenido histórico original y su procedencia no deben desaparecer. Las correcciones se añaden como nuevas capas de evidencia, no como sustitución retroactiva del hecho registrado.

La minimización de datos continúa siendo aplicable. Un RSC rico no justifica retener indefinidamente información

personal innecesaria ni convertir auditabilidad en vigilancia total. La arquitectura debe admitir referencias, pseudonimización, segregación de acceso, ventanas de retención por riesgo y preservación de metadatos suficientes cuando el contenido sensible deba eliminarse. La proporcionalidad entre evidencia y privacidad debe definirse por dominio.

La calidad del RSC condiciona toda inferencia posterior. Si falta el estado relevante, si las alternativas no son recuperables o si el impacto real nunca se mide, una afirmación de integración puede tener baja robustez. En ese caso, el sistema de control debe registrar explícitamente la clase de evidencia disponible y la incertidumbre. La ausencia de datos no debe llenarse con una reconstrucción narrativa no marcada.

4.4. Familia 3: gobernanza de la integración

La tercera familia es la contribución más específica de la arquitectura. Los marcos de monitoreo suelen registrar eventos y métricas de desempeño; aquí se exige además gobernar la decisión que convierte un evento en candidato a modificar el estado funcional. Para cada evaluación material deben conservarse: el evento de RSC examinado; la versión del evaluador; los principios, reglas o restricciones relevantes; la evidencia utilizada; la decisión de integrar o rechazar; la razón estructurada; el dominio afectado; el tipo de cambio esperado; la reversibilidad prevista; y las pruebas que deberán ejecutarse para comprobar el efecto.

El sistema debe distinguir al menos cuatro resultados de evaluación: confirmación de una regla existente, modificación de una regla o pauta, creación de un principio operativo nuevo y rechazo activo. La primera puede requerir vigilancia ligera; las dos intermedias son constitutivas y requieren mayor evidencia; el rechazo debe conservar la razón para evitar que la ausencia de integración sea interpretada retrospectivamente como olvido. Una quinta salida, evidencia insuficiente, bloquea la decisión constitutiva hasta disponer de información adecuada.

La integración no puede validarse por autodeclaración. Una entrada en $I(t)$ es una pretensión de estado que debe ser contrastable. El control exige una hipótesis de efecto: qué comportamiento debería cambiar, en qué población o clase de situaciones, en qué dirección y dentro de qué horizonte temporal. Posteriormente, un módulo de evaluación independiente compara el comportamiento observado con esa hipótesis y con explicaciones alternativas, entre ellas cambios de datos, versiones, contexto o intervenciones humanas.

La separación de funciones es especialmente importante. El componente que propone integrar un evento no debería ser la única autoridad que certifica que la integración ocurrió y produjo el efecto adecuado. Cuando el riesgo lo justifique, la arquitectura debe separar proponente, evaluador, aprobador y auditor. La separación puede ser organizacional, técnica o ambas. En dominios críticos, determinadas clases de integración —por ejemplo, cambios que afecten criterios de acceso a servicios esenciales— deberían requerir aprobación externa antes de influir en ϕ_1 .

También deben existir invariantes no modificables por integración autónoma. La segunda mediación no puede convertirse en una vía para eludir requisitos regulatorios, controles de seguridad, límites de propósito o derechos de las personas afectadas. Por ello, la EEA debe identificar restricciones superiores que no forman parte del espacio de aprendizaje constitutivo. Un intento de modificar esas restricciones se trata como evento de escalamiento y no como evolución ordinaria.

4.5. Familia 4: monitoreo longitudinal del estado y del efecto

El monitoreo longitudinal combina dos clases de señales. La primera procede del estado constitutivo: número y tipo de integraciones, dominios afectados, profundidad declarada, frecuencia, concentración de cambios, divergencia entre dominios y evolución de variables candidatas como S , α , Δ e IIC. La segunda procede del desempeño y del entorno:

métricas convencionales de calidad, drift, seguridad, sesgo, robustez, tasas de error, reclamaciones e incidentes. Ninguna clase sustituye a la otra.

El artículo adopta una regla de doble confirmación para cambios materiales. Una señal interna puede activar revisión, pero la existencia de un cambio constitutivo debe apoyarse en evidencia de estado y en efecto externo o conductual. De modo recíproco, un deterioro de desempeño no debe atribuirse automáticamente a $I(t)$ si existe una explicación más plausible en datos, infraestructura o contexto. El monitoreo debe mantener un modelo de hipótesis competidoras sobre la procedencia del cambio.

S puede actuar cualitativamente como alerta de creciente idiosincrasia; α como descripción operacional de cuánto se ha asumido o integrado bajo el protocolo de dominio; Δ como brecha entre atribución causal y asunción; e IIC como coherencia entre dominios o sub- $I(t)$. La congruencia entre principios declarados y conducta debe revisarse por separado mediante evidencia observacional. Sin calibración empírica, estos constructos no deben activar sanciones automáticas por un número fijo. Pueden, en cambio, priorizar auditorías, aumentar muestreo, reducir autonomía o exigir revisión humana cuando convergen con evidencia adicional.

La monitorización no debe esperar necesariamente a un incidente. WATCH demuestra que es posible diseñar métodos de monitoreo secuencial capaces de detectar changepoints y distinguir ciertos tipos de shift con control estadístico de falsas alarmas (Prinster et al., 2025). Una arquitectura de identidad acumulativa puede usar técnicas semejantes para el efecto observable, pero necesita además la CPC para atribuir la señal a una integración concreta. La novedad no está en detectar que algo cambió, sino en conservar la relación entre el cambio detectado y la historia de decisiones constitutivas que podría explicarlo.

La frecuencia de revisión debe ser proporcional al riesgo y a la velocidad de cambio. Un sistema con integraciones raras y dominio estable puede revisarse por lotes; un sistema con aprendizaje continuo en decisiones de alto impacto puede requerir monitoreo casi continuo de determinadas señales y revisión humana periódica. La frecuencia debe formar parte del expediente de gobernanza y no quedar implícita en herramientas operativas.

4.6. Familia 5: escalamiento, contención y revisión externa

El escalamiento es la respuesta a una pérdida de confianza suficientemente material, no un castigo al sistema. La arquitectura debe definir por anticipado qué eventos suspenden la operación ordinaria. Entre los disparadores razonables se encuentran: salida de la EEA; imposibilidad de reconstruir la CPC de un cambio material; divergencia persistente entre comportamiento declarado y observado; deterioro abrupto de desempeño o equidad; concentración anómala de integraciones en un dominio; conflictos entre dominios; divergencia sistemática entre autoevaluación y auditoría externa; e intento de modificar invariantes superiores.

La respuesta debe ser graduada. El primer nivel aumenta observabilidad o muestreo. El segundo reduce autonomía, alcance, población o throughput. El tercero congela nuevas integraciones y opera con una versión de estado previamente verificada. El cuarto suspende el uso material y activa revisión extraordinaria. El quinto, cuando el riesgo no puede ser mitigado con confianza, exige retirada, rediseño o nueva evaluación formal. Estos niveles no tienen por qué corresponder a una escala numérica universal; deben adaptarse al dominio y al régimen jurídico aplicable.

La contención debe ser técnicamente ejecutable. No basta una política que diga 'intervenir si algo sale mal'. Deben existir mecanismos para detener la actualización de $I(t)$, aislar dominios, revertir configuraciones derivadas cuando sea legítimo, preservar evidencia y asegurar que un rollback funcional no destruya el historial de eventos. La distinción RSC/ $I(t)$ es crítica: restaurar un estado previo no debe borrar el conocimiento de que los eventos intermedios ocurrieron.

La revisión externa cumple dos funciones: corregir sesgo autoservidor y aportar una perspectiva que el propio sistema no controla. El NIST AI RMF destaca el valor de la revisión independiente en determinadas actividades de medición y gestión; la literatura de certificación continua subraya la separación de funciones entre auditado, auditor y certificador (Knoblauch & Großmann, 2023). Para sistemas de segunda mediación, esa separación es aún más relevante cuando el sistema produce parte de la evidencia sobre su propio estado.

4.7. Familia 6: revisión extraordinaria de identidad funcional

La sexta familia responde a la pregunta más delicada: ¿cuándo la trayectoria acumulada hace que el sistema operativo sea materialmente distinto del objeto originalmente evaluado? Se propone una revisión extraordinaria de identidad funcional, que puede desembocar en nueva línea base, modificación de la EEA, contención o, cuando el régimen normativo lo exija, nueva evaluación de conformidad o recertificación. El término recertificación se usa aquí en sentido funcional y no pretende crear una certificación ISO autónoma.

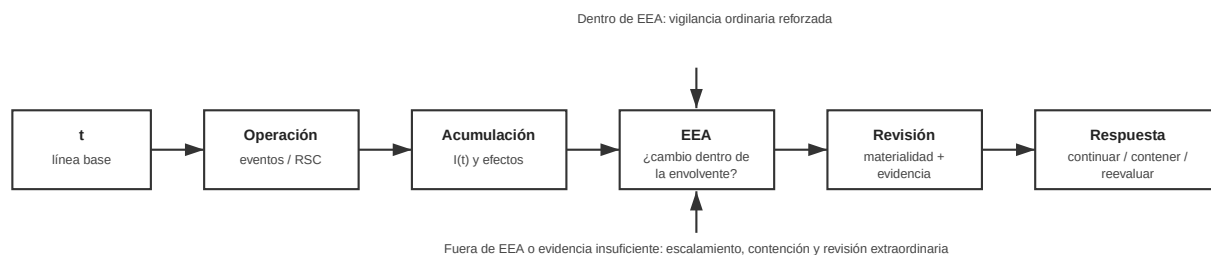
La revisión debe comparar t_0 con t_n en cinco dimensiones: propósito efectivo; dominios de operación; principios o restricciones funcionales; distribución de comportamiento relevante; y arquitectura/evidencia del mecanismo que produjo el cambio. Un cambio será más material cuando altere varias dimensiones, afecte población de alto impacto, se extienda entre dominios o reduzca la capacidad de reconstruir la procedencia del comportamiento.

La EEA funciona como presunción condicionada. Mientras el cambio esté previsto, documentado, medido y dentro de límites, puede mantenerse la equivalencia regulatoria práctica con la línea base. Fuera de la EEA, la equivalencia deja de presumirse. Esto no significa automáticamente que el sistema sea inseguro o ilícito; significa que la evidencia que justificó su autorización inicial ya no basta por sí sola.

El concepto dialoga directamente con el tratamiento de modificaciones sustanciales en el Reglamento europeo de IA. El artículo 43 exige una nueva evaluación de conformidad para sistemas de alto riesgo sujetos a modificación sustancial y aclara que, para sistemas que continúan aprendiendo, los cambios predeterminados por el proveedor en la evaluación inicial y documentados técnicamente no constituyen por ese solo hecho modificación sustancial. La EEA propone trasladar esa lógica a un nivel de control más explícito: no sólo predeterminar que habrá aprendizaje, sino especificar las clases de integración, los dominios, la evidencia y los límites dentro de los cuales dicho aprendizaje conserva la identidad funcional evaluada.

La revisión extraordinaria debe concluir con una decisión documentada: continuidad sin cambios; continuidad con controles reforzados; actualización de línea base; reducción de la EEA; ampliación de la EEA con nueva evidencia; suspensión parcial; o nueva evaluación externa. Esa decisión se convierte, a su vez, en parte de la historia de gobernanza del sistema.

Figura 2. Ciclo de revisión y envoltorio de evolución autorizada



Fuente: elaboración propia. EEA = envolvente de evolución autorizada.

4.8. Matriz operativa evento–riesgo–control–evidencia–respuesta

Tabla 2. Matriz operativa de controles

Evento / condición | Riesgo de gobernanza | Control | Evidencia mínima | Respuesta

Cambio técnico convencional | Pérdida de reproducibilidad o regresión. | Gestión de configuración + pruebas. | Versión, hash, datos, pruebas, aprobación. | Actualizar línea base si procede.

Evento material de operación | No poder reconstruir la decisión. | RSC reforzado. | Input/referencia, contexto, output, impacto. | Conservar y someter a evaluación.

Integración constitutiva propuesta | Cambio de reglas sin justificación. | Gobernanza de integración. | Evento, razón, principio afectado, efecto esperado. | Aprobar, rechazar o elevar.

Efecto no previsto tras integración | Pérdida de conformidad o comportamiento emergente. | Pruebas longitudinales + CPC. | Comparación t_0/t_n , drift, causalidad de cambio. | Contener y revisar.

Salida de EEA | Objeto operativo ya no cubierto por evaluación inicial. | Revisión extraordinaria. | Informe de materialidad + evidencia del cambio. | Nueva línea base / nueva evaluación / suspensión.

Divergencia autoevaluación–auditor | Sesgo autoservidor o fallo de observabilidad. | Verificación independiente. | Dos canales de evidencia y discrepancias. | Reducir autonomía y recalibrar.

Conflicto entre dominios | Fragmentación funcional. | Revisión IIC e invariantes transversales. | Estados por dominio, decisiones de transferencia. | Aislar dominios o escalar.

Evidencia insuficiente | Falsa precisión y atribución indebida. | Salida explícita de incertidumbre. | Mapa de datos faltantes y supuestos. | Bloquear cambio constitutivo material.

Fuente: elaboración propia.

V. Correspondencia normativa y de gestión

5.1. ISO/IEC 42001 y normas complementarias

ISO/IEC 42001:2023 especifica requisitos para establecer, implementar, mantener y mejorar continuamente un sistema de gestión de inteligencia artificial (SGIA). Su alcance organizacional y su énfasis en gestión de riesgos, trazabilidad, transparencia y mejora continua la convierten en el marco ISO más adecuado para situar la propuesta. La correspondencia que sigue es funcional y de alto nivel: no se afirma que las seis familias satisfagan por sí solas cláusulas específicas ni que la arquitectura constituya una certificación conforme a ISO/IEC 42001.

La línea base y la EEA se alinean conceptualmente con la necesidad de definir contexto, alcance, objetivos, responsabilidades y procesos del SGIA. La trazabilidad del RSC proporciona evidencia para operación, evaluación y mejora. La gobernanza de la integración añade un mecanismo específico de control de cambios funcionales. El monitoreo longitudinal y el escalamiento son compatibles con ciclos de evaluación del desempeño, tratamiento del riesgo, acciones correctivas y mejora continua. La revisión extraordinaria traduce esas capacidades al problema particular

de un sistema cuyo estado funcional puede evolucionar después del despliegue.

ISO/IEC 42005:2025 complementa esta arquitectura al proporcionar orientación para evaluaciones de impacto de sistemas de IA a lo largo del ciclo de vida, desde diseño y desarrollo hasta despliegue y monitoreo posmercado. En una arquitectura de identidad acumulativa, la evaluación de impacto no debería cerrarse en t_0 ; una modificación de $I(t)$ que amplíe el dominio, cambie una pauta constitutiva o altere poblaciones afectadas puede justificar una actualización de la evaluación.

ISO/IEC 23894:2023 aporta un marco de gestión de riesgos específico para IA y puede sustentar la proporcionalidad de controles y escalamiento. ISO/IEC 38507:2022 sitúa la IA en el plano de gobernanza organizacional y ayuda a distribuir responsabilidades entre órgano de gobierno, ejecutivos, especialistas, proveedores y auditores. Finalmente, ISO/IEC 42006:2025 establece requisitos adicionales para organismos que auditan y certifican SGIA conforme a ISO/IEC 42001. Su relevancia para este artículo se limita a la independencia y rigor del proceso de certificación; no debe confundirse la revisión de identidad funcional propuesta con una certificación ISO de producto.

5.2. NIST AI RMF

El NIST AI Risk Management Framework 1.0 organiza la gestión en cuatro funciones: GOVERN, MAP, MEASURE y MANAGE, concebidas para operar a lo largo del ciclo de vida (NIST, 2023). La arquitectura propuesta puede insertarse sin dificultad en esa lógica. GOVERN aloja las reglas de integración, las responsabilidades, invariantes y EEA; MAP caracteriza propósito, contexto, población, dominios y consecuencias; MEASURE conserva evidencia, observa el estado y contrasta efectos; MANAGE decide escalamiento, contención, corrección y revisión extraordinaria.

La correspondencia es útil porque evita presentar identidad acumulativa como un sistema de gobernanza aislado. El objeto novedoso puede gestionarse mediante funciones conocidas siempre que se añadan evidencias y decisiones específicas. Conviene señalar que, a septiembre de 2026, NIST ha anunciado que AI RMF 1.0 está en proceso de revisión. En consecuencia, la correspondencia aquí descrita se refiere a AI RMF 1.0 y deberá interpretarse a la luz de las revisiones posteriores que NIST publique.

5.3. Reglamento europeo de Inteligencia Artificial

El Reglamento (UE) 2024/1689 ofrece el contraste normativo más preciso. Su artículo 9 define la gestión de riesgos de sistemas de alto riesgo como un proceso continuo e iterativo planificado y ejecutado durante todo el ciclo de vida. El artículo 12 exige capacidades técnicas de registro automático de eventos durante la vida del sistema y vincula el logging con identificación de situaciones de riesgo, modificaciones sustanciales y monitoreo poscomercialización. El artículo 72 exige a proveedores de sistemas de alto riesgo establecer y documentar un sistema de monitoreo poscomercialización que recopile, documente y analice activamente datos de desempeño a lo largo de la vida del sistema.

Estos requisitos respaldan tres premisas de la arquitectura: la gobernanza es longitudinal, los eventos necesitan trazabilidad y la evidencia posdespliegue debe retroalimentar la gestión del riesgo. No obstante, el RSC propuesto es semánticamente más específico que un requisito general de logging porque busca preservar elementos que permitan reconstruir una decisión constitutiva y su efecto posterior. Esa especificidad debe verse como propuesta de control adicional y no como interpretación obligatoria del artículo 12.

El artículo 43 aporta el vínculo con la EEA. Los sistemas de alto riesgo que ya pasaron una evaluación de conformidad deben someterse a una nueva evaluación en caso de modificación sustancial. Para sistemas que continúan aprendiendo, determinados cambios predeterminados por el proveedor al momento de la evaluación inicial y documentados en la documentación técnica no constituyen modificación sustancial. La EEA generaliza esa intuición en lenguaje de

gobernanza: la evolución puede ser autorizada ex ante si se especifica su espacio, evidencia y límites; cuando la trayectoria excede esos límites, la equivalencia con el objeto inicialmente evaluado debe ser revisada.

El Reglamento (UE) 2026/1744, de 8 de julio de 2026 y publicado en el Diario Oficial el 24 de julio de 2026, modificó el calendario de aplicación del Reglamento (UE) 2024/1689. El capítulo III, secciones 1, 2 y 3 —con la excepción prevista en el propio artículo 113— será aplicable desde el 2 de diciembre de 2027 para sistemas de alto riesgo clasificados conforme al artículo 6(2) y el anexo III, y desde el 2 de agosto de 2028 para los clasificados conforme al artículo 6(1) y el anexo I. Por ello, el escenario de crédito de la sección VI debe leerse como análisis de un régimen vigente cuya fecha de aplicación para esas obligaciones específicas aún es futura.

5.4. OCDE y Consejo de Europa

Los Principios de IA de la OCDE, actualizados en 2024, exigen trazabilidad y un enfoque sistemático de gestión de riesgos durante el ciclo de vida. Su principio de robustez, seguridad y protección plantea que los riesgos deben evaluarse y gestionarse continuamente y que deben existir mecanismos para intervenir, reparar o retirar sistemas cuando aparezca comportamiento indebido. El principio de accountability vincula trazabilidad con datasets, procesos y decisiones. Estas formulaciones son compatibles con la CPC y con la exigencia de mantener evidencia sobre la procedencia de cambios relevantes.

El Consejo de Europa aporta una perspectiva de derechos. La Convención Marco sobre Inteligencia Artificial y Derechos Humanos, Democracia y Estado de Derecho, abierta a firma en 2024, adopta un enfoque de ciclo de vida y exige mecanismos de gestión de riesgos e impactos. La metodología HUDERIA fue aprobada por el Comité de Ministros el 26 de febrero de 2025 y el HUDERIA Model: Context-Based Risk Analysis (COBRA) fue aprobado el 25 de febrero de 2026. Ambos proporcionan orientación no vinculante y estructurada para evaluar contexto, diseño, despliegue, riesgos e impactos sobre derechos humanos, democracia y Estado de derecho. En sistemas de identidad acumulativa, esa evaluación debería poder actualizarse cuando la trayectoria funcional modifique materialmente el contexto de riesgo.

Tabla 3. Correspondencia funcional de alto nivel

Familia propia | ISO/IEC | NIST AI RMF | AI Act UE | Otros marcos

1. Línea base + EEA | 42001: sistema de gestión y mejora; 42005: impacto. | GOVERN / MAP | Documentación técnica; cambios predeterminados; art. 43. | OCDE: gestión del riesgo por ciclo de vida.
2. Trazabilidad/RSC | 42001: trazabilidad y control documental. | MAP / MEASURE | Art. 12; art. 72. | OCDE: trazabilidad; Consejo de Europa: documentación.
3. Integración | Control de cambios dentro del SGIA. | GOVERN / MEASURE | No existe categoría equivalente; control propuesto adicional. | Reviewability y auditoría como antecedentes.
4. Monitoreo longitudinal | 42005: actualización de impactos; 23894: riesgo. | MEASURE | Art. 9 y art. 72. | OCDE robustez; HUDERIA.
5. Escalamiento/contención | 23894; 38507. | MANAGE | Gestión del riesgo e incidentes. | OCDE: override/reparación/retiro.
6. Revisión identidad funcional | 42001: revisión/mejora; 42006: independencia del certificador. | GOVERN / MANAGE | Art. 43: nueva evaluación ante modificación sustancial. | Continuous assurance / CABC.

Nota: la tabla identifica correspondencias funcionales; no constituye una matriz de conformidad ni reproduce requisitos de normas comerciales.

VI. Escenario demostrativo: sistema de evaluación crediticia con integración longitudinal

Considérese un sistema hipotético de evaluación de solvencia de personas naturales desplegado por una entidad

financiera. El ejemplo se elige porque el anexo III del Reglamento europeo de IA clasifica como alto riesgo los sistemas destinados a evaluar la solvencia de personas naturales o establecer su puntuación crediticia, salvo sistemas destinados a detectar fraude financiero. No se pretende describir un sistema existente ni establecer una recomendación de crédito; el escenario sirve únicamente para mostrar la propuesta de control.

En t_0 , el proveedor establece y documenta la línea base técnica M_0 . El sistema utiliza datos financieros permitidos, un conjunto documentado de variables, una función φ_1 que produce una puntuación y una recomendación, límites de uso, umbrales de revisión humana y controles de no discriminación. $I(t_0)$ contiene únicamente principios iniciales autorizados. La EEA permite integraciones dentro del dominio 'tratamiento de incertidumbre documental', pero prohíbe que el sistema introduzca autónomamente nuevas variables sensibles, cambie la definición de finalidad, transfiera principios desde otros dominios o reduzca los requisitos de revisión humana para rechazos de alto impacto.

En t_1 se presenta una solicitud con ingresos irregulares pero un historial prolongado de cumplimiento. φ_1 produce una recomendación desfavorable con alta incertidumbre. Debido al impacto y a la divergencia entre la puntuación y el resultado de una revisión humana posterior, el evento activa un RSC reforzado. El registro conserva la versión de M_0 , las variables utilizadas, la distribución o intervalos de confianza disponibles, el output, la decisión humana, las razones documentadas y el resultado posterior del crédito cuando éste sea conocido bajo reglas de privacidad y retención.

En t_2 , el evaluador de segunda mediación analiza el evento junto con otros casos análogos. No concluye que el sistema 'fue injusto' como esencia moral; identifica una discrepancia operacional: la pauta de incertidumbre penaliza de manera sistemática determinada estructura de ingresos legítimos. La propuesta de integración consiste en modificar la política de manejo de incertidumbre para que, ante una combinación definida de señales, el sistema aumente la probabilidad de revisión humana en lugar de endurecer automáticamente el score. La integración se mantiene dentro de la EEA porque no cambia variables sensibles ni finalidad y porque su efecto previsto es aumentar revisión humana en una clase de casos previamente autorizada.

La decisión constitutiva genera una entrada de estado $I(t_2)$ vinculada a los eventos que la justifican, al principio operativo modificado, a su dominio, a la evidencia y a una hipótesis de efecto. La hipótesis puede formularse sin fijar un umbral universal: en casos análogos, la tasa de remisión a revisión humana debería aumentar y la tasa de falsos rechazos debería disminuir sin producir un deterioro material de morosidad o una disparidad nueva entre grupos protegidos. La arquitectura registra también qué resultados refutarían la eficacia de la integración.

Entre t_2 y t_3 , φ_1 continúa operando. El monitoreo compara la cohorte afectada con la línea base y con controles adecuados. Si la remisión aumenta como se esperaba, los falsos rechazos disminuyen y no aparecen riesgos adicionales, la CPC queda fortalecida: evento $\rightarrow$ evaluación $\rightarrow$ integración $\rightarrow$ modificación de estado $\rightarrow$ efecto observable. Si el comportamiento cambia en otra dirección, la integración no puede considerarse validada por el mero hecho de estar escrita en $I(t)$. Se activa revisión y puede revertirse el efecto funcional sin borrar el historial.

Supóngase ahora que, después de múltiples integraciones, el sistema comienza a extender el principio de revisión reforzada a solicitudes fuera del dominio autorizado y a modificar indirectamente el peso de variables que no estaban dentro de la EEA. La versión de software sigue llamándose M_0 y no existe un despliegue convencional de M_1 . Sin embargo, la distribución de decisiones ha cambiado de manera material y la CPC muestra que la trayectoria acumulada de integración es una causa plausible. El sistema ha salido de la EEA. La respuesta no es declarar automáticamente incumplimiento, sino suspender nuevas integraciones, limitar el dominio, conservar evidencia y activar revisión extraordinaria de identidad funcional.

La revisión compara t_0 y t_n . Si concluye que el comportamiento sigue dentro del propósito, que los riesgos permanecen

adecuadamente controlados y que la ampliación puede documentarse ex ante, puede establecerse una nueva línea base t_n y una EEA revisada. Si el cambio afecta la conformidad o el propósito en términos que encajen en la noción jurídica de modificación sustancial, el régimen aplicable puede exigir nueva evaluación de conformidad. La arquitectura de identidad acumulativa no sustituye esa decisión jurídica; produce la evidencia necesaria para formularla.

El ejemplo muestra por qué el problema no es equivalente a concept drift. Podría no haber cambiado la distribución de entrada. Tampoco es necesariamente una actualización MLOps: la misma versión técnica puede conservarse. Y no es simple monitoreo de desempeño: la gobernanza necesita saber qué decisión de integración produjo el cambio y si ese cambio estaba previamente autorizado. La CPC y la EEA ofrecen precisamente esos dos elementos.

Tabla 4. Secuencia del escenario crediticio

Momento	Estado / evento	Evidencia	Decisión de gobernanza
---------	-----------------	-----------	------------------------

t_0	$M_0 + I(t_0)$; EEA definida.	Baseline técnico y funcional.	Operación autorizada.
-------	--------------------------------	-------------------------------	-----------------------

t_1	Evento de crédito material.	RSC reforzado + revisión humana.	Enviar a evaluación ϕ_2 .
-------	-----------------------------	----------------------------------	--------------------------------

t_2	Integración dentro de EEA.	Razón, principio, efecto esperado.	Permitir cambio con monitoreo.
-------	----------------------------	------------------------------------	--------------------------------

t_3	Efecto medido.	Métricas de revisión, error, riesgo y equidad.	Confirmar, corregir o revertir.
-------	----------------	--	---------------------------------

t_n	Trayectoria excede EEA.	CPC acumulada + comparación t_0/t_n .	Contener y revisar identidad funcional.
-------	-------------------------	---	---

Fuente: escenario hipotético de elaboración propia.

VII. Discusión y límites

La arquitectura propuesta debe evaluarse por la precisión del problema que resuelve, no por la amplitud de sus metáforas. Su primera ventaja es separar objetos que suelen aparecer mezclados: logs, estados, cambios, aprendizaje y cumplimiento. El RSC preserva historia; $I(t)$ representa un estado derivado; la CPC conecta ese estado con efectos; y la EEA delimita qué evolución permanece cubierta por una autorización previa. Esa separación permite incorporar sistemas adaptativos a marcos de gobernanza existentes sin exigir que toda modificación sea tratada como un sistema nuevo.

La segunda ventaja es que la propuesta es falsable en un sentido operacional. Si un sistema declara que integró un evento pero no existe modificación conductual cuando debería observarse, la hipótesis de integración queda debilitada. Si existe cambio sin una CPC suficiente, la gobernanza no puede atribuirlo a segunda mediación y debe mantener hipótesis alternativas. Si las salidas de la EEA no correlacionan con cambios materiales, el diseño de la envolvente requiere recalibración. La arquitectura no necesita asumir que el sistema siempre conoce correctamente su propio estado.

El primer límite es empírico. El Volumen XI formula una arquitectura y perfiles de implementación, pero reconoce que la calibración de varios elementos requiere datos. Este artículo no aporta un experimento que demuestre una implementación completa de ϕ_2 ni valida umbrales de S , α , Δ o IIC. En particular, los valores numéricos históricos no deben trasladarse a políticas como si fueran límites universales. La siguiente etapa de investigación debe construir implementaciones mínimas y estudiar si los eventos, decisiones de integración y efectos posteriores pueden identificarse con suficiente fiabilidad.

El segundo límite es de observabilidad. Sistemas cerrados, modelos de terceros, arquitecturas distribuidas o agentes con herramientas externas pueden impedir el acceso a v_t , alternativas o decisiones internas. La arquitectura admite perfiles de evidencia parcial, pero la reducción de observabilidad disminuye la fuerza de las afirmaciones sobre procedencia del cambio. Una inferencia retrospectiva desde outputs y logs puede ser útil, pero no equivale a acceso directo a ϕ_2 . La calidad de la gobernanza debe hacer visible esa degradación epistémica.

El tercer límite es causal. R^* y la propia CPC dependen de atribuir causas en sistemas complejos. Un cambio observado puede tener múltiples mecanismos y las rutas pueden interactuar de forma no aditiva. La evidencia temporal no basta para causalidad. En entornos de alto riesgo, la atribución debería combinar experimentos controlados cuando sean posibles, análisis contrafactual, comparación de cohortes, pruebas de regresión y revisión humana. El artículo no propone un algoritmo causal universal.

El cuarto límite es organizacional. Un sistema técnicamente instrumentado puede seguir siendo mal gobernado si incentivos, presiones comerciales o estructuras de responsabilidad favorecen omitir eventos, rebajar severidad o aceptar autoevaluaciones convenientes. La literatura de accountability ha mostrado que los problemas algorítmicos son sociotécnicos (Wieringa, 2020; Raji et al., 2020; Cobbe & Singh, 2020). Por ello, CPC y EEA deben insertarse en asignaciones reales de autoridad, separación de funciones, auditoría y canales de impugnación.

El quinto límite es normativo. La correspondencia con ISO/IEC es conceptual; no se ha reproducido ni auditado el texto íntegro de las normas comerciales. El AI Act, por su parte, contiene conceptos jurídicos cuya aplicación depende del caso, de guías y de autoridades competentes. La EEA no es sinónimo de cambio predeterminado bajo el artículo 43 y no reemplaza la noción jurídica de modificación sustancial. Su función es producir una disciplina documental que puede ayudar a responder esas preguntas, no resolverlas por definición.

El sexto límite es semántico. La palabra identidad puede invitar a lecturas antropomórficas. Por ello, el artículo utiliza identidad funcional o identidad causal acumulativa de forma estrictamente operacional. No se infiere conciencia, intención moral, dignidad, voluntad jurídica ni capacidad para ser titular de derechos. La proposición regulatoria es más modesta: si una trayectoria interna documentada modifica de forma persistente el comportamiento, esa trayectoria es relevante para determinar qué objeto se está gobernando.

La discusión con Glukhov (2026) permite precisar aún más el punto. Si la identidad de un modelo depende de convenciones de equivalencia, la gobernanza debe elegir criterios adecuados a su propósito. La EEA puede entenderse como una convención explícita de equivalencia regulatoria: define qué cambios siguen perteneciendo al mismo objeto evaluado para fines de control. Esa convención debe ser revisable, transparente y vinculada al riesgo, no una afirmación ontológica absoluta.

Finalmente, el trabajo de Lahiri (2026) sobre reconfiguración sociotécnica iniciada por agentes sugiere una extensión futura. La arquitectura actual gobierna cambios funcionales derivados de integración; una generación posterior debería estudiar sistemas capaces de modificar también sus propios canales de supervisión, frecuencia de monitoreo o estructura de coordinación. En ese escenario, la EEA tendría que cubrir no sólo el comportamiento de tarea, sino también la metagobernanza: quién puede cambiar las reglas mediante las cuales se decide qué cambio está autorizado.

VIII. Conclusiones

Los marcos contemporáneos de IA ya reconocen que los sistemas cambian, que el riesgo debe gestionarse durante el ciclo de vida y que el monitoreo posdespliegue es indispensable. Por esa razón, la contribución de este artículo no puede formularse como una simple crítica a una certificación estática. El problema específico aparece cuando la procedencia del cambio incluye una trayectoria selectiva de eventos evaluados e integrados en un estado funcional que condiciona decisiones futuras.

Para gobernar ese supuesto se ha propuesto una arquitectura de seis familias de control. La línea base identifica el objeto y la EEA delimita su evolución autorizada. El RSC preserva la historia de eventos. La gobernanza de la integración documenta por qué un evento modifica o no el estado derivado. El monitoreo longitudinal contrasta estado y comportamiento. El escalamiento reduce confianza y autonomía cuando la evidencia se deteriora o la trayectoria excede

límites. La revisión extraordinaria decide si la equivalencia con el objeto inicialmente evaluado puede mantenerse, debe actualizarse o requiere una nueva evaluación externa.

Dos distinciones sostienen el modelo. La primera es historia frente a estado: RSC e I(t) no son el mismo objeto. La segunda es estado frente a evidencia: una declaración de integración no es una observación directa de ϕ_2 y necesita una cadena probatoria que conecte evento, evaluación, decisión, estado y efecto posterior. La CPC evita atribuir a identidad acumulativa todo cambio observado; la EEA evita que aprendizaje autorizado se confunda con modificación material por el solo hecho de ser cambio.

La propuesta es compatible, en términos funcionales, con ISO/IEC 42001, ISO/IEC 42005, NIST AI RMF, el enfoque de ciclo de vida de la OCDE, HUDERIA y varios elementos del AI Act europeo. Especialmente relevante es la distinción europea entre cambios predeterminados de sistemas que continúan aprendiendo y modificaciones sustanciales que requieren nueva evaluación. La arquitectura propuesta no sustituye esos marcos: aporta una semántica adicional de procedencia y materialidad para sistemas con estado funcional acumulativo.

S, α , Δ e IIC cumplen en esta propuesta funciones analíticas y de priorización de revisión. Su utilización como métricas de decisión o umbrales normativos requiere operacionalización, calibración y validación específicas para cada dominio. La evaluación empírica de esas operacionalizaciones constituye una línea posterior de trabajo, distinta de la arquitectura de gobernanza aquí formulada.

La respuesta a la pregunta principal puede formularse finalmente en una regla: cuando un sistema puede cambiar porque integra su propia historia, gobernar el sistema exige gobernar también la historia que puede llegar a constituirlo, la decisión que selecciona esa historia y la evidencia que demuestra qué cambió después.

Referencias

- Ashmore, R., Calinescu, R., & Paterson, C. (2021). Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Computing Surveys*, 54(5), Article 111, 1–39. <https://doi.org/10.1145/3453444>
- Balendran, A., Benchoufi, M., Evgeniou, T., & Ravaud, P. (2024). Algorithmovigilance, lessons from pharmacovigilance. *npj Digital Medicine*, 7, 270. <https://doi.org/10.1038/s41746-024-01237-y>
- Calinescu, R., Weyns, D., Gerasimou, S., Iftikhar, M. U., Habli, I., & Kelly, T. (2018). Engineering trustworthy self-adaptive software with dynamic assurance cases. *IEEE Transactions on Software Engineering*, 44(11), 1039–1069. <https://doi.org/10.1109/TSE.2017.2738640>
- Cobbe, J., & Singh, J. (2020). Reviewable automated decision-making. *Computer Law & Security Review*, 39, 105475. <https://doi.org/10.1016/j.clsr.2020.105475>
- Council of Europe. (2024). Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225).
- Council of Europe. (2025). HUDERIA Methodology — Risk and impact assessment of AI systems from the point of view of human rights, democracy and the rule of law. <https://www.coe.int/en/web/artificial-intelligence/huderia-risk-and-impact-assessment-of-ai-systems>
- Council of Europe. (2026). HUDERIA Model: Context-Based Risk Analysis (COBRA). Approved 25 February 2026. <https://www.coe.int/en/web/artificial-intelligence/huderia-risk-and-impact-assessment-of-ai-systems>
- De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., & Tuytelaars, T. (2022). A

- continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7), 3366–3385. <https://doi.org/10.1109/TPAMI.2021.3057446>
- European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L 2024/1689.
- European Parliament & Council of the European Union. (2026). Regulation (EU) 2026/1744 of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). *Official Journal of the European Union*, L 2026/1744, 24 July 2026. <https://eur-lex.europa.eu/eli/reg/2026/1744/oj>
- Foalem, P. L., Da Silva, L., Khomh, F., Li, H., & Merlo, E. (2025). Logging requirement for continuous auditing of responsible machine learning-based applications. *Empirical Software Engineering*, 30, Article 97. <https://doi.org/10.1007/s10664-025-10656-8>
- Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), Article 44, 1–37. <https://doi.org/10.1145/2523813>
- Glukhov, V. (2026). Position: Model identity in machine learning is a convention, not a property. Accepted at the ICML 2026 Position Track. SSRN. <https://doi.org/10.2139/ssrn.7078918>
- International Organization for Standardization & International Electrotechnical Commission. (2022). ISO/IEC 38507:2022 Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations.
- International Organization for Standardization & International Electrotechnical Commission. (2023a). ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management.
- International Organization for Standardization & International Electrotechnical Commission. (2023b). ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system.
- International Organization for Standardization & International Electrotechnical Commission. (2025a). ISO/IEC 42005:2025 Information technology — Artificial intelligence (AI) — AI system impact assessment. <https://www.iso.org/standard/42005>
- International Organization for Standardization & International Electrotechnical Commission. (2025b). ISO/IEC 42006:2025 Information technology — Artificial intelligence — Requirements for bodies providing audit and certification of artificial intelligence management systems. <https://www.iso.org/standard/42006>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Knoblauch, D., & Großmann, J. (2023). Towards a risk-based continuous auditing-based certification for machine learning. *The Review of Socionetwork Strategies*, 17, 255–273. <https://doi.org/10.1007/s12626-023-00148-w>
- Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine learning operations (MLOps): Overview, definition, and architecture. *IEEE Access*, 11, 31866–31879. <https://doi.org/10.1109/ACCESS.2023.3262138>
- Lahiri, R. (2026). Agent-initiated socio-technical reconfiguration: A three-level taxonomy of autonomous AI

governance and its recursive challenges. *Frontiers in Artificial Intelligence*, 9, 1881783. <https://doi.org/10.3389/frai.2026.1881783>

Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & Elish, M. C. (2021). Algorithmic impact assessments and accountability: The co-construction of impacts. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 735–746. <https://doi.org/10.1145/3442188.3445935>

Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021). Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics*, 27, 44. <https://doi.org/10.1007/s11948-021-00319-4>

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, 26, 2141–2168. <https://doi.org/10.1007/s11948-019-00165-5>

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). <https://doi.org/10.6028/NIST.AI.100-1>

OECD. (2024). OECD AI Principles: Recommendation of the Council on Artificial Intelligence (updated May 2024). Organisation for Economic Co-operation and Development.

Olvera Aguilar, M. H. (2026). La integración del ego en una IA (Vol. XI). Meriadock.

Prinster, D., Han, X., Liu, A., & Saria, S. (2025). WATCH: Adaptive monitoring for AI deployments via weighted-conformal martingales. *Proceedings of the 42nd International Conference on Machine Learning, Proceedings of Machine Learning Research*, 267, 49830–49859.

Rabanser, S., Günnemann, S., & Lipton, Z. C. (2019). Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, 32, 1396–1408.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. <https://doi.org/10.1145/3351095.3372873>

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28, 2503–2511.

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 1–18. <https://doi.org/10.1145/3351095.3372833>