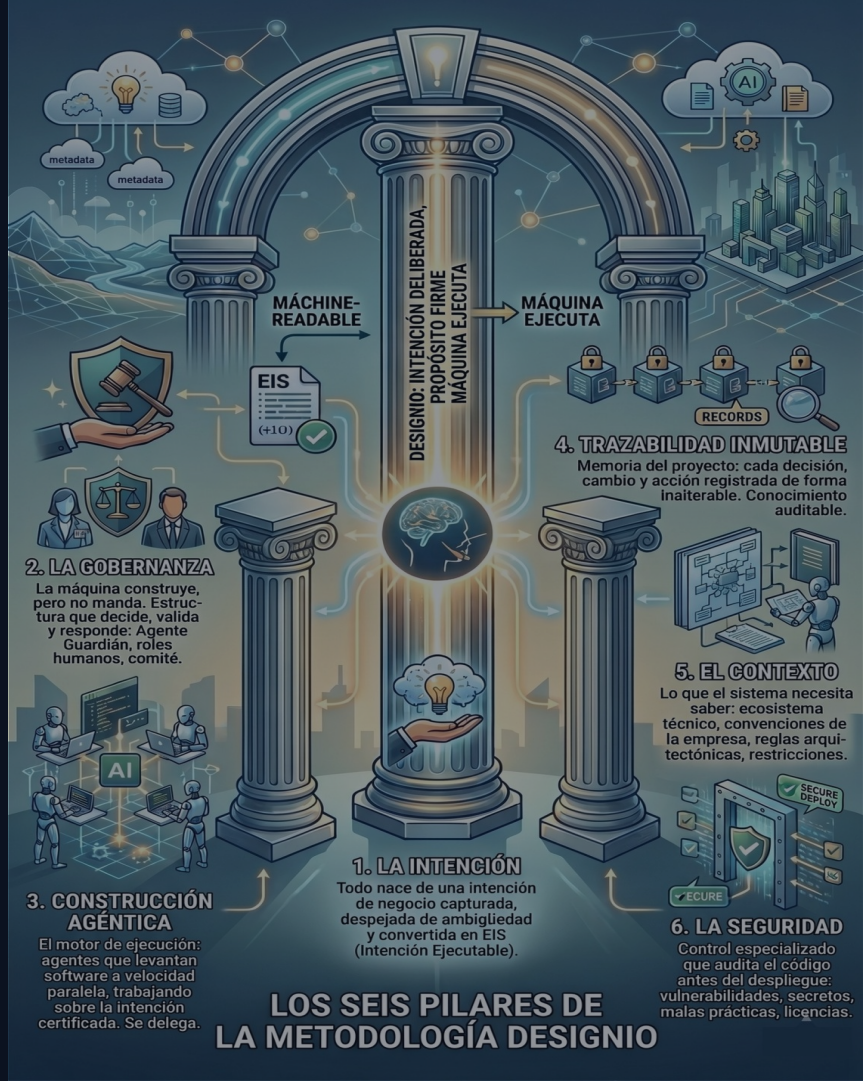
PROPUESTA DE SEGURIDAD · PARA EL CISO

Soberanía de la Intención

Seguridad por diseño para construir software con agentes de IA — y por qué Designio frente a cualquier otra alternativa.

Metodología Designio · Marcos Fernández Otero



CONTEXTO

2026: el punto de ruptura agéntico

La IA ya escribe miles de líneas de código al día en tu organización. La pregunta dejó de ser “¿usamos agentes?”. Ahora es:

¿Cómo aceleramos con IA sin entregar el control de nuestro propio código?

Velocidad extrema

Los agentes generan y despliegan en horas lo que antes llevaba semanas.

Autonomía sin gobierno

Sin reglas, un agente decide y actúa sobre tu infraestructura solo.

Trazabilidad perdida

Nadie sabe qué requisito originó cada línea, ni quién la aprobó.

EL PROBLEMA

La “Caja Negra”: acelerar sin gobernanza

Código mezclado de humano y agente, aprobado sin leerse, sin vínculo a un requisito concreto, con añadidos de otros agentes sin contexto. El resultado es una caja negra más difícil de abrir que la de un avión.

- Código de agente aprobado por un humano que no tuvo tiempo de leerlo
- Líneas no trazadas ni vinculadas a ningún requisito
- Capas de distintos agentes y desarrolladores sin contexto compartido

> ***“¿Acelerarías en un coche sin volante? Eso es construir con IA sin gobernanza.”***

EL RIESGO

Qué pasa cuando no hay seguridad

Fuga de IP y credenciales

Una sola filtración de propiedad intelectual puede comprometer el proyecto entero.

Acciones destructivas del agente

Sin límites, un agente puede ejecutar acciones irreversibles (el caso “OpenClaw”).

Pérdida de soberanía

Nadie en la empresa entiende ya su propio código: dependes de la caja negra.

Incumplimiento normativo

Sin trazabilidad ni control de acceso, la auditoría y el cumplimiento son imposibles.

Coste de incidente

Detectar, contener y reconstruir tras un fallo no gobernado es caro e impredecible.

Reescritura masiva

Un error de diseño detectado tarde obliga a rehacerlo casi todo.

LA SOLUCIÓN

Designio en una frase

Gobernanza + intención firmada + agentes fortificados.
El humano decide el QUÉ; el agente ejecuta el CÓMO, dentro de barreras.

Gobernanza

AIGB, umbrales de confianza y reglas de enganche definen qué puede hacer la IA.

Intención firmada

Cada cambio nace de un EIS validado: trazabilidad de extremo a extremo.

Contención

Entorno aislado, egress controlado e identidad efímera por agente.

CADENA DE CUSTODIA

EIS: la intención firmada que ancla cada línea

La EIS —Especificación Ejecutable de Intenciones (Executable Intent Specification)— es el contrato validado del que nace el código. Convierte la caja negra en una cadena de custodia auditable.

Resultado: ningún cambio existe sin un requisito firmado que lo respalde. Auditoría inmediata.

Seguridad por capas: el blindaje de Designio

Ejecución aislada	Micro-VMs efímeras; el agente nunca toca la máquina anfitriona
Red y egress	Proxy + lista blanca + DPI + umbral de volumen
Identidad	Identidad sintética por agente, JWT de vida corta
Secretos	Inyección con mínimo privilegio, nunca credenciales compartidas
Observabilidad	Cada acción registrada y trazada al EIS
Validación	Revisión cruzada + certificación de seguridad

Cada capa es independiente: si una falla, las demás contienen el daño.

CAPA DE RED

Aislamiento y control de salida (egress)

Los agentes no tienen acceso libre a internet. Cada conexión saliente pasa por un proxy que valida tres cosas antes de dejar salir el tráfico:

¿A dónde va?

Solo destinos en la lista blanca del sprint: endpoints de modelos, repos privados y servicios que la EIS necesita.

¿Qué lleva dentro?

Inspección profunda (DPI) del contenido para detectar fugas de código sensible o credenciales.

¿Cuánto sale?

Si el volumen supera los umbrales habituales, el sistema pausa y exige verificación humana.

Un destino fuera de la lista blanca se registra como anomalía: se congela la instancia o se bloquea la conexión.

CAPA DE IDENTIDAD

Identidad sintética y tokens efímeros

Cada agente que actúa es una entidad independiente: no comparte credenciales ni usa la cuenta del proyecto. Se gestiona como un usuario humano de alto privilegio, con un token JWT propio.

Vida corta

El token caduca al terminar el sprint; no queda activo de forma indefinida.

Mínimo privilegio

Solo los accesos estrictamente necesarios para la tarea concreta.

Trazabilidad total

Cada acción con ese token queda registrada y vinculada al EIS que la originó.

Sin credenciales compartidas no hay punto único de robo: el blast radius de un agente comprometido es mínimo.

GOBERNANZA

AI Governance Board y Matriz de Umbrales de Confianza

El AIGB define el marco; la Matriz de Umbrales de Confianza fija qué autonomía tiene la IA. Sin esta matriz el agente podría hacer demasiado — o cosas que jamás debería hacer.

NIVEL 1 · AUTONOMÍA TOTAL

Formateo, orden de imports, eliminación de código muerto y warnings triviales.

NIVEL 2 · REQUIERE ACUERDO

Cambios de lógica, dependencias o esquema: necesitan validación humana.

NIVEL 3 · PROHIBIDO

Acciones críticas o irreversibles sobre datos, pagos o infraestructura.

Rules of Engagement + protocolo de resolución de conflictos humano-IA cierran el marco.

CONTENCIÓN

El “fuerte”: Agente Guardián, Whitelists y Railguards

Un Agente Guardián intercepta cada instrucción antes de que impacte en la infraestructura del cliente. Los agentes constructores nunca tocan la shell directamente.

Proxy de comandos

Toda shell o terminal pasa por el Guardián.

Whitelist de terminal

Solo comandos explícitamente permitidos se ejecutan.

Barrera financiera

Railguards frenan desviaciones de presupuesto y tokens.

VERIFICACIÓN

Observabilidad, revisión cruzada y certificación

Observabilidad agéntica

Cada acción del agente queda registrada y es reproducible. Si algo se desvía, se ve en tiempo real.

Pipeline con revisión cruzada

Ningún cambio llega a producción sin pasar por revisión y por una certificación de seguridad en Pre.

Endurecimiento final

Pruebas de estrés y pentesting adversarial antes de operar con un cliente nuevo.

Construir CON vs SIN Designio

DIMENSIÓN	SIN DESIGNIO	CON DESIGNIO
Trazabilidad	Caja negra: nadie sabe el origen	Cada línea ligada a un EIS firmado
Control de egress	Salida libre, riesgo de fuga	Proxy + lista blanca + DPI
Identidad	Cuentas y credenciales compartidas	Identidad sintética + JWT efímero
Autonomía del agente	Sin límites (riesgo “OpenClaw”)	Matriz de umbrales + Agente Guardián
Gobernanza	Inexistente o informal	AIGB + Rules of Engagement
Auditoría / cumplimiento	Imposible de demostrar	Trazabilidad total y certificación
Coste de incidentes	Alto e impredecible	Contenido y predecible

EL COSTE DE NO HACERLO

Anatomía de un incidente sin gobernanza

Con Designio, cada paso tiene un freno

Entorno aislado, egress controlado, identidad efímera, Agente Guardián y trazabilidad al EIS contienen el daño antes de que se propague — y permiten reconstruir exactamente qué ocurrió.

CUMPLIMIENTO

Alineado con los marcos que te importan

Los controles de Designio se alinean con los estándares de seguridad de IA que tu equipo ya debe demostrar:

CONTROL DESIGNIO	RIESGO QUE MITIGA	MARCO DE REFERENCIA
Egress + DPI + whitelist	Fuga de información sensible	OWASP LLM06 · ISO 27001 A.8
Identidad sintética + mínimo privilegio	Acceso indebido / robo de credenciales	ISO 27001 A.5 · NIST AI RMF (Govern)
Matriz de umbrales + Agente Guardián	Exceso de agencia del modelo	OWASP LLM08 · EU AI Act (supervisión)
Trazabilidad al EIS	Falta de registro y auditabilidad	EU AI Act (trazabilidad) · ISO 42001
Revisión cruzada + certificación	Cadena de suministro insegura	OWASP LLM05 · NIST (Measure)
Observabilidad agéntica	Incidentes no detectados	NIST AI RMF (Manage) · SOC 2

Alineamiento de controles, no certificación formal: el alcance del endurecimiento se ajusta a tu sector y regulación.

MODELO DE RESPONSABILIDAD

Qué aporta el cliente y cómo arrancamos seguros

EL CLIENTE APORTA

Identidad

Proveedor de identidad para las identidades sintéticas

Conectividad

Acceso controlado a repos y servicios necesarios

Secretos

Bóveda de secretos con inyección de mínimo privilegio

Telemetría de retorno

Canal para métricas y trazas del sprint

CHECKLIST DE READINESS

- [x] Pruebas de humo de conectividad — Bloque 1
- [x] Auditoría de la línea base semántica — Bloque 2
- [x] Gobernanza y firmas en regla — Bloque 3
- [x] Sprint cero (dry run) completo — Bloque 4

EL NEGOCIO

El dividendo de la intención: seguridad rentable

La seguridad de Designio no es un coste: es lo que hace que la IA sea rentable de verdad. Gobernar la intención elimina retrabajo y convierte el gasto en activo.

Fin del “Impuesto a la Ambigüedad”

El EIS elimina el retrabajo por requisitos mal entendidos.

Tokenomics vs Headcount

El coste se mide en tokens gobernados, no en plantilla creciente.

Software como activo soberano

Código trazable, auditable y propiedad real de la empresa.

ADOPCIÓN

Hoja de ruta: de cero a producción gobernada

01

Fase 0 · Cimientos

Diagnóstico de madurez, AIGB, matriz de umbrales, system prompts y whitelists.

02

Readiness

Checklist de conectividad, identidad y secretos. Gobernanza y firmas listas.

03

Sprint cero

Dry run completo del ciclo en un entorno seguro antes de tocar producción.

04

Sprint productivo

Ciclo semanal de 7 fases con certificación de seguridad en cada release.

OBJECIONES DEL CISO

Las preguntas difíciles, respondidas

“¿La IA tendrá acceso a nuestros secretos?”

No. Identidad sintética por agente, JWT efímero y secretos inyectados con mínimo privilegio. Nunca credenciales compartidas.

“¿Cómo evito que un agente haga algo destructivo?”

La Matriz de Umbrales y el Agente Guardián interceptan cada comando. Lo crítico exige acuerdo humano explícito.

“¿Y si filtra nuestro código?”

Egress por proxy + lista blanca + DPI + umbral de volumen. Cualquier salida anómala pausa el sistema.

“¿Podré auditar lo que hizo la IA?”

Sí. Trazabilidad total: cada acción queda ligada a un EIS firmado, con observabilidad en tiempo real.

No frenamos la IA. La gobernamos.

Designio convierte el desarrollo con agentes en un proceso seguro, trazable y auditable — y transforma tu software en un activo soberano.

Siguiente paso: diagnóstico de madurez del SDLC + un sprint cero (dry run) para ver el blindaje en acción.