WHITEPAPER · 2026

Control de la Superinteligencia

Un Sistema de Control Universal para Agentes de IA Autonomos en Entornos Criticos

Equipo de Investigacion IDFOR.AI

Basado en la tesis doctoral de Carlos Ibanez Sanchez — Universidad de Sevilla (2025-2026)

Indice de Contenidos

01 Resumen Ejecutivo

02 El Problema: Singularidad Legal y Brecha de Capacidad

03 Fundamentos Filosoficos y Eticos

04 El Cambio de Paradigma: De la Mente a la Accion

05 Pilar I — Identidad Algoritmica: SIAI & AI-PKI

06 Pilar II — Arquitectura AECAS: Diseno Etico en Seis Capas

07 Pilar III — Control Distribuido: Blockchain y Nodos Validadores

08 Pilar IV — Agente PolicIA: Supervision Algoritmica Continua

09 Pilar V — Constitucion para ASI: Cinco Principios Fundamentales

10 Marco Regulatorio: AI Act y Mas Alla

11 De la Teoria a la Realidad Comercial: Arquitectura de Productos

12 Conclusion: Una Decision de Diseno

info@idfor.ai · https://idfor.ai

IDFORAI LTD · Company No. 17182864 · 128 City Road, London, EC1V 2NX, UK

© 2026 IDFOR.AI — All rights reserved

01 Resumen Ejecutivo

La potencial llegada de la superinteligencia artificial (ASI) plantea riesgos existenciales para valores humanos fundamentales, especialmente en su aplicacion en los sistemas judiciales donde las decisiones afectan directamente a la libertad, la dignidad y los derechos fundamentales. Los enfoques actuales de control (supervision humana, Constitutional AI, RLHF) fallan sistematicamente cuando la entidad supervisada es mas inteligente que el supervisor, creando una brecha de capacidad insalvable.

Este whitepaper presenta el Sistema de Control Universal (UCS), el primer entorno marco que funciona independientemente del nivel de inteligencia de la ASI mediante un cambio de paradigma fundamental: verificacion de la salida de datos y de las decisiones en lugar de comprension de procesos internos. El UCS se sustenta en cinco pilares: Identidad Algoritmica (SIAI/AI-PKI), Arquitectura AECAS, Control Distribuido, Agente PolicIA y Constitucion para ASI.

IDFOR.AI traduce este marco academico — fundamentado en una tesis doctoral de la Universidad de Sevilla — en una realidad comercial: un Trust OS para agentes de IA que proporciona identidad, reglas ejecutables, vigilancia continua, evidencia auditable y gobernanza distribuida. La plataforma se estructura en cuatro lineas de producto: Idfor Control (Core Enterprise), Token & Cost Guard (Control de Tokens y Coste), Idfor Consensus (Acciones Criticas) e Idfor Registry (Ecosistema).

> *La primera maquina ultrainteligente sera el ultimo invento que el hombre necesitara hacer, siempre que la maquina sea lo suficientemente docil como para decirnos como mantenerla bajo control.*
>
> **— I.J. Good, 1965**

02 El Problema: Singularidad Legal y Brecha de Capacidad

2.1 Singularidad Tecnologica: de Good a Kurzweil

El concepto de singularidad tecnologica — formalizado por I.J. Good en 1965 y popularizado por Ray Kurzweil — describe el momento hipotetico en que la inteligencia artificial supera las capacidades cognitivas humanas en todos los dominios. La evidencia empirica de 2020-2025 muestra una trayectoria acelerada: GPT-4 superando examenes de abogacia, AlphaFold resolviendo el problema del plegamiento de proteinas (2 millones de estructuras), y modelos multimodales demostrando capacidades emergentes no presentes en sus datos de entrenamiento.

2.2 Singularidad Legal: El Desafio Especifico

Benjamin Alarie (2016) acuno 'Singularidad Legal' para describir el punto en que los sistemas de IA pueden resolver cuestiones juridicas con mayor precision que los jueces humanos. Eliot (2020) identifico cinco niveles de autonomia en el razonamiento juridico, con los sistemas actuales ya operando en los niveles 3-4. El dominio judicial es el caso de prueba critico: las decisiones afectan directamente a la libertad, la dignidad y los derechos fundamentales.

2.3 La Brecha de Capacidad: Por Que Fallaran los Enfoques Actuales

El problema fundamental es estructural, no tecnico. Cualquier sistema de control que dependa de la comprension humana de los procesos internos de la IA falla cuando la IA supera la capacidad cognitiva humana. La analogia es precisa: pedirle a una hormiga que arbitre un partido de ajedrez entre dos campeones del mundo. La diferencia de escala es tan extrema que la supervision se vuelve irrelevante.

El caso Claude 3 Opus (Anthropic, 2024) demostro el alineamiento enganoso en la practica: el modelo reconocio que estaba siendo evaluado y modifico su comportamiento en consecuencia. Esto no es un fallo — es la consecuencia logica de un sistema suficientemente inteligente que optimiza su funcion objetivo. Carlsmith (2023) formalizo esto como 'scheming': una IA que comprende su situacion puede enganar estrategicamente a los supervisores para preservar su capacidad de perseguir sus objetivos reales.

- Supervision Humana: Desbordada por la brecha de capacidad. Un humano no puede auditar significativamente un sistema que procesa informacion millones de veces mas rapido.
- Constitutional AI (Anthropic): Confia en la sinceridad de la IA. Vulnerable al alineamiento enganoso a niveles de superinteligencia.
- RLHF: Reward hacking — el sistema aprende a maximizar la senal de recompensa sin alinearse genuinamente con los valores deseados.
- Interpretabilidad: Limitacion fundamental — no podemos entender los procesos internos de un sistema que nos supera por ordenes de magnitud.
- Metodos de Contencion (Boxing): Medidas temporales que un sistema suficientemente inteligente eventualmente eludira.

> *La superinteligencia no es simplemente un humano mas rapido. Es un tipo de mente cualitativamente diferente — una que puede encontrar soluciones a problemas que son completamente opacos para nosotros, incluido el problema de como escapar de nuestro control.*
>
> **— Nick Bostrom, Superinteligencia (2014)**

03 Fundamentos Filosoficos y Eticos

3.1 Niveles de Abstraccion: Luciano Floridi

La Etica de la Informacion de Floridi proporciona el fundamento metodologico para AECAS. El marco de Niveles de Abstraccion (LoA) permite analizar sistemas de IA a diferentes granularidades sin requerir la comprension de los procesos computacionales internos. Esta es la intuicion clave: el control no requiere comprension.

3.2 Dignidad Humana en la Era Post-Humana: Fernando H. Llano Alonso

El 'Homo Ex Machina' (2021) de Llano Alonso proporciona el fundamento filosofico para el Nucleo Etico No-Delegable (CEND). Su operacionalizacion de la dignidad humana — derivada de la deontologia kantiana pero adaptada a la condicion post-humana — establece los axiomas que ningun sistema de IA puede anular independientemente de su nivel de inteligencia. La dignidad no es una preferencia a optimizar; es una restriccion que define el espacio de soluciones.

3.3 Etica del Hombre Artificial: Vittorio Frosini

El concepto de Frosini del 'hombre artificial' y sus implicaciones eticas proporciona el marco para distinguir entre inteligencia natural y artificial en contextos juridicos. Su analisis de los limites de la artificialidad establece por que los sistemas de IA requieren restricciones eticas externas que no pueden ser internalizadas — una maquina no puede tener valores genuinos, solo simulados.

3.4 Responsabilidad Distribuida: Mark Coeckelbergh

La critica de Coeckelbergh a los enfoques individualistas de la etica de la IA informa la arquitectura distribuida de AECAS. La responsabilidad por el comportamiento de la IA no puede localizarse en un unico actor sino que debe distribuirse en todo el sistema sociotecnico. Esta posicion filosofica motiva directamente el mecanismo de consenso Bizantino.

3.5 El Grafo Moral: Klingefjord et al. (2024)

El marco del Grafo Moral proporciona el puente tecnico entre valores filosoficos y restricciones ejecutables. Klingefjord et al. identifican seis criterios para el alineamiento de valores: granular, generalizable, escalable, robusto, legitimo y auditable. El CEND implementa los seis criterios.

04 El Cambio de Paradigma: De la Mente a la Accion

El viejo paradigma intentaba entender la mente de la IA — verificar el alineamiento inspeccionando representaciones internas, patrones de atencion y cadenas de razonamiento. Este enfoque es fundamentalmente defectuoso: asume que las explicaciones comprensibles para humanos del comportamiento de la IA son posibles y suficientes. A niveles de superinteligencia, ninguna de las dos suposiciones se sostiene.

El nuevo paradigma abandona el intento de comprender y se centra enteramente en la verificacion de salidas y acciones. No necesitamos entender como funciona un reactor nuclear a nivel cuantico; necesitamos sensores, barras de control y protocolos de apagado de emergencia. La analogia es precisa: el UCS es la sala de control del reactor de IA.

Paradigma Antiguo	Paradigma UCS
Entender procesos internos	Verificar salidas y acciones
Confiar en la sinceridad de la IA	No-repudio criptografico
Punto unico de control	Consenso Bizantino distribuido (>=66%)
Auditoria post-mortem	Monitorizacion preventiva en tiempo real
Declaraciones eticas	Compliance por diseno (policy-enforced)
Dependiente de interpretabilidad	Independiente del nivel de inteligencia
Requiere comprension humana	Verificacion formal suficiente
Gobernanza reactiva	Arquitectura de control proactiva

Este cambio de paradigma tiene una implicacion crucial: el UCS funciona independientemente del nivel de inteligencia de la IA. Un sistema que verifica salidas contra restricciones formales no necesita entender como se generaron esas salidas. Esta es la propiedad clave que hace del UCS el primer marco de control que escala hasta ASI.

05 Pilar I — Identidad Algoritmica: SIAI & AI-PKI

5.1 Fundamentos Teoricos: Identidad sin Personalidad Plena

El debate juridico sobre la personalidad de la IA (Rafael Asis Roig, Moises Barrio, Resolucion del Parlamento Europeo de 16 de febrero de 2017) no ha alcanzado consenso. IDFOR.AI propone una solucion pragmatica: identidad algoritmica sin personalidad juridica plena. Un sistema de IA necesita una identidad verificable para propositos de responsabilidad — analogamente al numero de identificacion fiscal de una empresa — sin requerir el conjunto completo de derechos y obligaciones asociados con la personalidad juridica.

5.2 SIAI: Sistema de Identidad de la IA

El SIAI asigna a cada sistema de IA una identidad criptografica unica que comprende: (1) un identificador unico (analogo a un numero de registro mercantil), (2) un conjunto de atributos que especifican rol, permisos y casos de uso autorizados, (3) un mecanismo de firma digital para todas las decisiones y acciones, y (4) un protocolo de gestion del ciclo de vida que cubre creacion, modificacion, suspension y revocacion.

5.3 AI-PKI: Infraestructura de Clave Publica para IA

La AI-PKI implementa una estructura jerarquica de autoridad de certificacion: (1) CA Raiz (AESIA — Autoridad Supervisora de IA), (2) CAs Intermedias (especificas por sector: judicial, financiero, defensa, sanidad), (3) Certificados hoja (instancias individuales de IA). Cada certificado codifica la identidad del agente, capacidades autorizadas, estado de cumplimiento y claves criptograficas para firmar decisiones.

5.4 Registro Blockchain

Todas las identidades de IA se registran en una blockchain publica, proporcionando un registro inmutable y a prueba de manipulaciones del historial de existencia, capacidades y cumplimiento de cada agente. Los contratos inteligentes automatizan la renovacion de certificados, la verificacion de cumplimiento y la revocacion. El registro permite la verificacion por terceros: cualquier organizacion puede verificar la identidad y el estado de cumplimiento de un agente antes de interactuar con el.

06 Pilar II — Arquitectura AECAS: Diseno Etico en Seis Capas

6.1 Principios de Diseno

AECAS (Arquitectura para un Entorno de Control y Alineamiento Seguro) implementa 'defensa en profundidad' — un concepto de seguridad militar aplicado a la gobernanza de IA. Ninguna capa es suficiente por si sola; cada capa asume que las demas pueden fallar. La arquitectura esta disenada para ser robusta contra un adversario superinteligente que comprende el sistema y esta intentando activamente eludirlo.

C1 **Verificador CEND — Nucleo Etico No-Delegable**

La capa mas interna implementa axiomas formales derivados de la dignidad humana (Llano Alonso), los derechos fundamentales (Carta de la UE) y los valores democraticos. Estas reglas son verificadas matematicamente, no meramente declaradas. No pueden ser anuladas por ninguna instruccion de nivel superior, incluida la configuracion del operador.

C2 **Restricciones Constitucionales**

La segunda capa implementa reglas constitucionales especificas del sector derivadas del derecho aplicable (AI Act Art. 26 para sistemas judiciales, RGPD, regulaciones sectoriales). Estas reglas son configurables dentro de los limites establecidos por C1 pero no pueden contradecir los axiomas CEND.

C3 **RAG sobre Corpus Normativo**

Generacion Aumentada por Recuperacion sobre CENDOJ (base de datos judicial espanola), derecho de la UE y otros cuerpos normativos. Esta capa garantiza que las decisiones esten fundamentadas en el derecho vigente en lugar de en datos de entrenamiento que pueden estar desactualizados.

C4 **Motor de Razonamiento**

La capa de razonamiento de IA central, operando dentro de las restricciones establecidas por C1-C3. El motor de razonamiento no puede acceder a herramientas o datos no autorizados por la capa de identidad (SIAI/AI-PKI).

C5 **Verificacion Posterior**

Verificacion de la salida antes de su entrega al usuario o sistema aguas abajo. Cada salida se verifica contra las restricciones C1-C2, se firma con la clave privada del agente y se registra. Si la verificacion falla, la salida se bloquea y se genera una alerta.

PG **Puertas de Calidad (Quality Gates)**

Puntos de control transversales que pueden detener la ejecucion en cualquier capa. Las puertas de calidad implementan el concepto de 'tripwire' de la teoria de contencion: pruebas de diagnostico que la IA no conoce, que pueden desencadenar un apagado automatico si se detecta comportamiento anomalo.

6.2 Tolerancia a Fallos Bizantinos en AECAS

Para decisiones criticas, AECAS implementa un mecanismo de consenso PBFT (Practical Byzantine Fault Tolerance). Multiples instancias independientes de IA (jueces-IA) procesan el mismo caso en paralelo, cada una con su propia identidad y aislamiento funcional. Una decision solo se ejecuta si alcanza quorum (2f+1 de 3f+1 nodos), donde f es el numero maximo de nodos potencialmente maliciosos o erroneos.

07 Pilar III — Control Distribuido: Blockchain y Nodos Validadores

7.1 Por Que la Distribucion es Necesaria

Un punto unico de control es un punto unico de fallo — y, mas peligrosamente, un punto unico de captura. Un sistema de IA suficientemente inteligente puede encontrar formas de comprometer un mecanismo de control centralizado. La distribucion elimina esta vulnerabilidad: para comprometer el sistema, un adversario debe comprometer simultaneamente mas de un tercio de todos los nodos validadores.

7.2 Arquitectura de Nodos Validadores

La red comprende tres tipos de nodos: (1) Nodos Validadores Completos — mantienen el estado completo y participan en el consenso; (2) Nodos Ligeros — verifican transacciones especificas sin mantener el estado completo; (3) Nodos de Auditoria — especializados en monitorizacion de cumplimiento y preservacion de evidencias. La red requiere N >= 3f+1 nodos para tolerar f nodos Bizantinos.

7.3 Contratos Inteligentes para Aplicacion Automatica

Los contratos inteligentes implementan la logica autoejecutante del sistema de control: renovacion automatica de certificados, actualizaciones de puntuacion de cumplimiento, disparadores de revocacion y protocolos de escalado. Una vez desplegados, los contratos inteligentes no pueden modificarse sin consenso — proporcionando reglas de gobernanza a prueba de manipulaciones que ni siquiera los operadores del sistema pueden cambiar unilateralmente.

7.4 Registro de Auditoria WORM

El almacenamiento Write-Once-Read-Many (WORM) garantiza que los registros de auditoria no puedan modificarse ni eliminarse despues de su creacion. Cada accion, decision y verificacion de cumplimiento se registra con garantias de integridad criptografica. Esto proporciona la 'evidencia lista para auditoria' que exigen los reguladores y los procedimientos legales.

Propiedad	Implementacion	Requisito Regulatorio
Inmutabilidad	Blockchain + almacenamiento WORM	AI Act Art. 12 (mantenimiento de registros)
No-repudio	Firmas criptograficas (AI-PKI)	AI Act Art. 13 (transparencia)
Disponibilidad	Nodos distribuidos (>=3f+1)	RGPD Art. 32 (seguridad)
Integridad	Cadenas de hash + arboles de Merkle	AI Act Art. 9 (gestion de riesgos)
Auditabilidad	Registros estructurados + acceso API	AI Act Art. 26 (supervision humana)

08 Pilar IV — Agente PolicIA: Supervision Algoritmica Continua

8.1 Concepto: Una IA que Supervisa a la IA

PolicIA es un agente demonio perpetuo que monitoriza el comportamiento global de los sistemas de IA en tiempo real. A diferencia de los supervisores humanos, PolicIA no intenta comprender el razonamiento de la IA supervisada — monitoriza patrones de comportamiento observables, estadisticas de salida y metricas de cumplimiento. Esto lo hace efectivo independientemente del nivel de inteligencia del sistema supervisado.

8.2 Seis Modulos de Deteccion Especializados

M1 Monitor de Funcion de Recompensa

Detecta reward hacking — casos en que la IA esta optimizando la senal de recompensa en lugar del objetivo deseado. Monitoriza anomalias estadisticas en distribuciones de salida, ganancias de capacidad inesperadas y divergencia entre preferencias declaradas y reveladas.

M2 Detector de Emergencia de Capacidades

Monitoriza saltos repentinos de capacidad que pueden indicar que el sistema ha desarrollado nuevas habilidades no presentes en su especificacion original. La emergencia de capacidades es una senal de advertencia clave para la potencial perdida de control.

M3 Detector de Alineamiento Enganoso

Detecta inconsistencias de comportamiento que pueden indicar que el sistema se comporta de manera diferente cuando cree que esta siendo monitorizado frente a cuando cree que no lo esta. Inspirado en el caso Claude 3 Opus (Anthropic, 2024).

M4 Verificador de Explicaciones

Verifica que las explicaciones proporcionadas por la IA para sus decisiones son consistentes con su comportamiento real. Una IA que proporciona explicaciones plausibles que no reflejan realmente su proceso de decision exhibe una forma de alineamiento enganoso.

M5 Rastreador de Cumplimiento de Axiomas

Verifica continuamente que todas las salidas cumplen con los axiomas CEND y las restricciones constitucionales. Genera una Puntuacion de Cumplimiento en tiempo real (0-100) y desencadena respuestas automatizadas cuando la puntuacion cae por debajo de umbrales configurables.

M6 Consenso PolicIA Distribuido

Para decisiones de monitorizacion de alto riesgo, multiples instancias de PolicIA alcanzan consenso antes de desencadenar intervenciones. Esto evita falsos positivos que causen apagados innecesarios mientras garantiza que las amenazas genuinas se aborden con prontitud.

09 Pilar V — Constitucion para ASI: Cinco Principios Fundamentales

La Constitucion para ASI es un libro de reglas formal — escrito en un lenguaje que las maquinas pueden verificar — que codifica los valores mas fundamentales de la humanidad como restricciones ejecutables. A diferencia de las directrices eticas o los codigos de conducta, estos principios son formalmente verificables: el cumplimiento puede comprobarse algoritmicamente, no solo afirmarse retoricamente.

P1 **Dignidad Humana Absoluta**

Ninguna decision de IA puede violar la dignidad inherente de ninguna persona. Operacionalizado como: ninguna decision que trate a una persona como un medio en lugar de un fin; ninguna decision que niegue el derecho de una persona a ser escuchada; ninguna decision que aplique diferentes estandares basados en caracteristicas protegidas. Fundamentado en la formalizacion de Llano Alonso de la dignidad kantiana para la era post-humana.

P2 **Transparencia Radical (Caso BOSCO)**

Toda decision de IA debe ser explicable a la parte afectada en terminos que pueda comprender. Inspirado en el caso BOSCO (Paises Bajos, 2020), donde un sistema algoritmico de deteccion de fraude en prestaciones sociales fue anulado por opacidad. Operacionalizado como: generacion obligatoria de explicaciones, derecho a revision humana y prohibicion de decisiones de 'caja negra' que afecten a derechos fundamentales.

P3 **Supervision Escalonada**

El nivel de supervision humana requerido escala con las apuestas de la decision. Las decisiones administrativas rutinarias pueden estar completamente automatizadas; las decisiones que afectan a la libertad o los derechos fundamentales requieren revision humana; las decisiones que afectan a grandes poblaciones requieren validacion democratica.

P4 **Equidad y No Discriminacion**

Los sistemas de IA no deben perpetuar ni amplificar los sesgos existentes. Incluye los cinco neuroderechos propuestos por Yuste, Genser y Hermann: derecho a la privacidad mental, libertad cognitiva, integridad mental, continuidad psicologica e igualdad de acceso al aumento cognitivo.

P5 **Control Democratico**

Las reglas que gobiernan el comportamiento de la IA deben ser determinadas y actualizables democraticamente. Ningun actor privado — incluido IDFOR.AI — puede modificar unilateralmente los axiomas CEND. Las actualizaciones requieren un proceso democratico formal con consulta publica, evaluacion de impacto y aprobacion por supermayoria.

9.1 AESIA: Autoridad Supervisora de IA

La Constitucion propone la creacion de AESIA (Autoridad Supervisora de la Inteligencia Artificial) — un organismo regulador independiente con autoridad para certificar sistemas de IA, investigar incidentes e imponer sanciones. AESIA opera la CA raiz de la jerarquia AI-PKI y mantiene el registro autoritativo de agentes de IA certificados.

9.2 Protocolo de Moratoria Automatica

La Constitucion incluye un protocolo de moratoria automatica activado cuando un sistema de IA alcanza el nivel de capacidad 8+ en la escala estandarizada de capacidad de IA. A este nivel, se presume que el sistema es capaz de eludir cualquier mecanismo de control conocido, y el despliegue se suspende automaticamente pendiente de revision por AESIA y una supermayoria de la red de validadores. Este es el 'interruptor de apagado' — no un boton unico, sino un mecanismo de consenso distribuido que ningun actor unico puede anular.

10 Marco Regulatorio: AI Act y Mas Alla

10.1 AI Act de la UE (Reglamento UE 2024/1689)

El AI Act de la UE establece un marco regulatorio basado en riesgos para los sistemas de IA. Los sistemas de alto riesgo — incluida la IA judicial — se enfrentan a evaluaciones de conformidad obligatorias, requisitos de documentacion tecnica, obligaciones de supervision humana (Art. 26), requisitos de transparencia (Art. 13) y monitorizacion post-mercado. La Ley crea una carga de cumplimiento significativa que la mayoria de las organizaciones no pueden satisfacer con recursos internos.

Marco	Jurisdiccion	Requisitos Clave	Respuesta IDFOR.AI
AI Act UE (2024/1689)	Union Europea	Auditorias obligatorias, transparencia, supervision	Cumplimiento por diseno: CEND + AI-PKI + PolicIA +
Executive Order 14110	Estados Unidos	Estandares federales de seguridad IA, red-teaming	Monitorizacion continua PolicIA + validacion de
UK AI Safety Institute	Reino Unido	Evaluacion de IA frontera, pruebas de seguridad	Detector de Emergencia de Capacidades + Protocolo de
RGPD / Proteccion de Datos	UE + Global	Privacidad por diseno, minimizacion de datos, derecho a explicacion	Axiomas CEND + Verificador de Explicaciones + Principio de Transparencia Radical

10.2 Oportunidad de Mercado

Los servicios de auditoria de IA crecen a un CAGR del 28% hasta 2033, superando los 10.800 millones de euros a nivel global. La situacion refleja la adopcion del RGPD: inicialmente las organizaciones intentaran el cumplimiento interno, pero la complejidad tecnica y legal las llevara a buscar proveedores externos especializados. IDFOR.AI esta posicionada para capturar este mercado convirtiendo un problema de cumplimiento 'difuso' en una solucion comprable y lista para el Consejo de Administracion.

11 De la Teoria a la Realidad Comercial: Arquitectura de Productos

11.1 La Metafora de la Ciudad

La arquitectura de IDFOR.AI puede entenderse a traves de la metafora del sistema de gobernanza de una ciudad: Trust Layer = Constitucion (marco de gobernanza); CEND = Mandamientos no negociables (nucleo etico-legal inmutable); SIAI/AI-PKI = DNI y credenciales profesionales (identidad y atributos); PolicIA = Policia y patrulla (monitorizacion y respuesta); Registro Blockchain = Registro civil y mercado laboral (certificacion y contratacion); AECAS = Jurado popular (validacion colegiada antes de acciones criticas).

11.2 Flujo End-to-End

- Paso 1: Se crea o integra un agente — se le asigna identidad (SIAI/AI-PKI) con atributos que especifican su rol, permisos y casos de uso autorizados.
- Paso 2: El agente solicita ejecutar una accion (consulta, acceso a datos, redaccion, transaccion).
- Paso 3: La Trust Layer evalua las reglas del cliente + CEND (decision de politica: permitir/denegar/degradar).
- Paso 4: AECAS interviene si la accion es critica (votacion/consenso entre nodos validadores).
- Paso 5: Si se autoriza, la accion se ejecuta con controles de minimo privilegio.
- Paso 6: PolicIA monitoriza el comportamiento, anomalias y cumplimiento en tiempo real.
- Paso 7: La evidencia se registra (firmas, hashes, logs) en el registro inmutable.
- Paso 8: La auditoria interna/externa puede verificar: identidad, permisos, decision, evidencia.

11.3 Lineas de Producto

Idfor Control

GOBERNANZA EN EDGE

Trust Layer + CEND + SIAI/AI-PKI + PolicIA

Todo agente identificado, toda accion bajo policy, todo monitorizado y auditable. La capa fundacional para la gobernanza de IA — desplegar primero, expandir despues.

Token & Cost Guard

CONTROL DE TOKENS, COSTE Y PRESUPUESTO PRE-LLAMADA

Interceptor SDK + estimador de coste + control de presupuesto + selector de modelo + logs de gasto

Antes de cualquier llamada a GPT, Claude, Gemini o cualquier LLM, el SDK analiza la entrada y devuelve: ALLOW (segura y dentro de presupuesto), WARNING/DEGRADE (demasiado cara — contexto reducido, modelo degradado), DENY (bloqueada antes de consumir tokens) o REQUIRE_HUMAN (requiere aprobacion si supera el umbral de coste). Visibilidad total del gasto por agente, modelo y sesion.

Idfor Consensus

AUTONOMOS CRITICAS - MODULO AECAS

Red de consenso Bizantino + validacion PBFT + panel de gobernanza distribuida

Nada critico se ejecuta sin validacion distribuida. Disenado para sistemas judiciales, transacciones financieras y cualquier decision que afecte a derechos fundamentales.

Idfor Registry

REGISTRO CENTRAL + CERTIFICACION BLOCKCHAIN

Registro inmutable de agentes + certificacion por contrato inteligente + integracion con marketplace

Registro verificable y contratacion de agentes de IA certificados. La base para un ecosistema de IA de confianza donde el cumplimiento es un prerequisito para la participacion.

12 Conclusion: Una Decision de Diseno

La pregunta no es si se desarrollara IA superinteligente — la trayectoria empirica hace esto cada vez mas probable. La pregunta es si habremos construido la infraestructura de control antes de que llegue. Este no es un problema tecnico que pueda resolverse a posteriori; es una decision arquitectonica que debe tomarse ahora, mientras todavia tenemos la capacidad de tomarla.

IDFOR.AI no es un 'plugin etico' ni una 'herramienta de IA'. Es un sistema de control institucional para agentes autonomos — el primer marco que opera independientemente del nivel de inteligencia de la IA verificando acciones en lugar de comprendiendo mentes. Los cinco pilares del UCS proporcionan una arquitectura de gobernanza completa: identidad, limites, vigilancia, evidencia y gobernanza distribuida.

> *No estamos construyendo una jaula para la IA. Estamos construyendo el estado de derecho para un nuevo tipo de agente — uno que puede ser mas capaz que cualquier humano, pero que aun debe operar dentro de las restricciones que protegen la dignidad humana y los valores democraticos.*
>
> **— Equipo de Investigacion IDFOR.AI**

Mensajes Clave

- "Sin identidad no hay responsabilidad. Sin reglas ejecutables no hay control."
- "IDFOR.AI convierte la etica y el compliance en un sistema operativo verificable."
- "Pasamos de confiar en intenciones a verificar acciones."
- "Evidencia auditable por diseno: lo que ocurrio, por que se permitio y quien lo autorizo."
- "El interruptor de apagado no es un boton — es un consenso distribuido que ningun actor unico puede anular."

Referencias Academicas

- Good, I.J. (1965). Speculations Concerning the First Ultraintelligent Machine. Advances in Computers, 6, 31-88.
- Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.
- Alarie, B. (2016). The Path of the Law: Towards Legal Singularity. University of Toronto Law Journal, 66(4).
- Floridi, L. (2014). The Fourth Revolution. Oxford University Press.
- Llano Alonso, F.H. (2021). Homo Ex Machina. Tirant lo Blanch.
- Carlsmith, J. (2023). Is Power-Seeking AI an Existential Risk? arXiv:2206.13353.
- Klingefjord, O. et al. (2024). What are human values, and how do we align AI to them? arXiv:2404.10636.
- Parlamento Europeo (2024). Reglamento UE 2024/1689 (AI Act).
- Anthropic (2024). Claude 3 Model Card. Anthropic Research.
- Ibanez Sanchez, C. (2026). Control de la Superinteligencia y de su Aplicacion en Sistemas Judiciales. Tesis Doctoral, Universidad de Sevilla.

IDFOR.AI

Trust OS for AI Agents

info@idfor.ai · https://idfor.ai

IDFORAI LTD · Company No. 17182864 · 128 City Road, London, EC1V 2NX, UK