CREDENCE · MOTOR DE DECISIÓN DETERMINISTA Y AUDITABLE

El LLM nunca decide.

Credence — la capa de decisión determinista y auditable para los agentes de IA de su entidad.

- MISMO EXPEDIENTE → MISMA DECISIÓN
- CADA PASO EXPLICADO
- CADA DECISIÓN RE-EJECUTABLE
- NADA SIN FIRMA HUMANA

EL MOMENTO

Los agentes ya trabajan en el banco. Lo que no pueden es responder por una decisión.

- **La categoría explota:** soporte, análisis documental, código y redacción ya corren en producción con agentes — la fiabilidad es hoy el principal cuello de botella del despliegue de IA (Stanford AI100, 2026).
- **12,3 %** de los encargos bancarios completados de forma autónoma por los mejores modelos en evaluación (ATLAS-Finance, State of AI Report 2026).
- En un fallo documentado, un agente frontier **aprobó sus 13 autocontroles** validando contra su propia hipótesis financiera errónea.
- Las explicaciones de los modelos **no son fieles** al proceso de decisión real; incidentes con agentes se han descubierto meses después, revisando transcripts a mano.

«La pregunta del comité de riesgos no es si el agente es capaz.
Es quién responde cuando decide mal.»

La tesis de esta propuesta: confinar el lenguaje al lenguaje. El agente percibe, redacta y explica; **la decisión es aritmética** — reproducible, explicada y congelada con su evidencia.

Kill switch, safety cases, contención por agente: la disciplina que los propios laboratorios se aplican antes de soltar un modelo — **convertida en producto** para los agentes del banco.

EL PROBLEMA

Un LLM en el camino de decisión no es certificable

Comportamiento de un agente LLM decidiendo		Lo que exige una entidad supervisada
Alucina cifras, hechos y acciones que nunca ocurrieron	→	**Reproducibilidad total:** mismo expediente → misma decisión
«El modelo decidió» no es defendible ante auditoría o supervisión	→	**Explicación de primera clase:** creencias, reglas y valores, no texto generado
Auditoría y replay imposibles; trazas que superan la supervisión humana	→	**Evidencia inmutable** y re-ejecución exacta de cualquier decisión pasada
Cada decisión consume tokens y depende de un proveedor externo	→	**Coste marginal ≈ 0**, milisegundos, en la infraestructura del banco
Sin garantía: solo texto plausible, sin óptimo demostrable	→	**Cada opción valorada en euros** — la decisión es un cálculo verificable

LA SOLUCIÓN

Credence: la decisión es un cálculo — no un prompt

La conducta del agente es el resultado de **aritmética sobre el vocabulario del negocio**: señales, creencias y costes. El LLM queda fuera del camino de decisión.

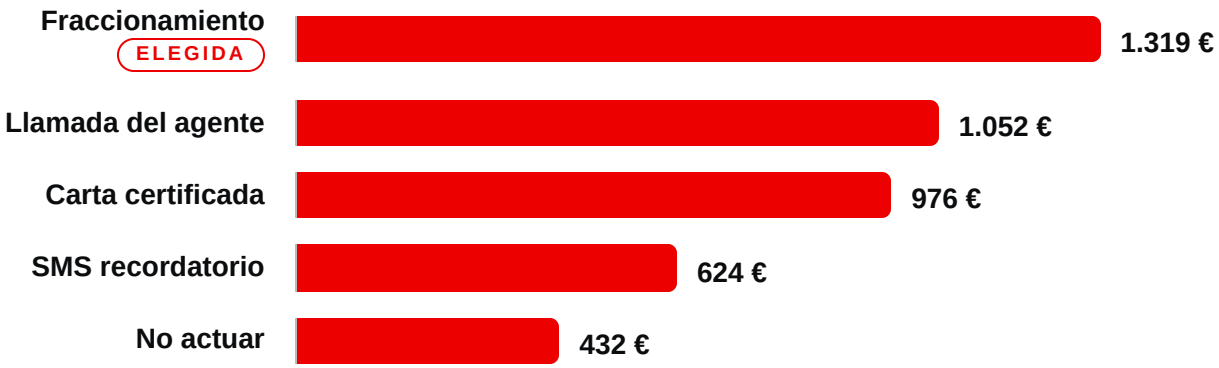
1	2	3	4	5	6	7
Señales	**Creencias**	**Acciones**	**Valor**	**Registro**	**Firma**	**Outcome**
evidencia del expediente, con la razón de cada dato	probabilidades con origen declarado — nunca certeza	todo lo posible hacer, con su coste real en euros	valor esperado de cada vía, en euros	decisión congelada con toda su evidencia	una persona aprueba — no hay modo automático	el resultado real recalibra el pack, bajo veto

Determinismo sellado. Mismo input → misma salida, hoy y dentro de un año. Cada decisión congela su configuración (hash), su semilla y la versión del motor.

Packs de decisión como datos. El vocabulario del dominio —acciones, costes, probabilidades, reglas— vive en base de datos: se edita desde el admin web, versionado, sin despliegues. Ajustar una política es un cambio de datos auditado, no un release.

UNA DECISIÓN BANCARIA, ABIERTA AL AUDITOR

Recobro: el valor de cada acción, en euros



Valor esperado neto por acción = P(pago) × recuperación esperada − coste de la acción. Ejemplo ilustrativo del pack de recobro; costes y probabilidades son filas versionadas, editables desde el admin.

Acción	Coste	P(pago)	Si éxito	Valor neto
Ofrecer fraccionamiento (3 pagos)	3,00 €	58 %	2.280 €	**1.319 €**
Llamada del agente	4,20 €	44 %	2.400 €	1.052 €
Carta certificada	8,50 €	41 %	2.400 €	976 €
SMS recordatorio	0,05 €	26 %	2.400 €	624 €
No actuar esta semana	0 €	18 %	2.400 €	432 €

EXPEDIENTE

Préstamo personal de **2.400 €** · 45 días de impago · 2 incidentes previos · ingresos verificados estables

CREENCIA USADA

P(pago espontáneo en 7 días) = **18 %**

origen declarado: prior de cartera vs. outcomes verificados de la red

Lo que emerge del cálculo: «fraccionar antes que escalar presión» no lo programó nadie — sale de la aritmética. La diferencia de **887 €** entre fraccionar y no actuar es el valor de la decisión. 12 ms por decisión · replay bit a bit desde el minuto uno.

POR QUÉ CREDENCE

Garantías estructurales — no aspiracionales

	Agente LLM	ML «black box»	Reglas duras (BRMS)	Credence
Mismo input → misma salida	No	Variable	Sí	**Garantizado**
Explicación	Inventada o ausente	Post-hoc, débil	El árbol IF	**Creencias, reglas y valor de cada vía**
Auditoría y replay	Imposible	Difícil	Limitado	**Inmutable + drift detectado**
Se adapta al caso	No	No	No — umbrales fijos	**Valora cada opción, cada vez**
Calibración con resultados	—	Reentrenar (opaco)	Manual	**Outcomes reales, bajo veto humano**
Coste y latencia	Tokens · segundos	Atado a proveedor	Rápido · rígido	**ms · ≈0 on-prem**

Un motor de reglas clásico es determinista — pero no valora alternativas ni aprende de resultados. Un modelo estadístico aprende — pero no es certificable. Credence es la intersección: cálculo determinista con probabilidades calibradas y costes reales.

EL ACTIVO

El registro de decisión: cada decisión nace con su evidencia

Snapshot congelado
Señales y hechos derivados, con la razón de cada uno, en el instante de proponer.

Identidad del cálculo
Hash de configuración del pack, versión del motor y semilla — replay exacto.

Creencias con origen
Cada probabilidad declara de dónde viene: prior público o outcomes verificados.

Actor y motivo
Quién aprobó, rechazó o ejecutó cada transición — append-only, cadena de hash.

Todas las alternativas
El valor de cada vía — no solo la ganadora. Se ve lo que se descartó y por qué.

Outcome observado
La decisión guarda su propio resultado: pagó, asistió, reclamó. El expediente se autoalimenta.

«El registro es inmutable y se re-ejecuta bit a bit.
El expediente se defiende solo.»

Ante una reclamación de cliente, auditoría interna o supervisión: reconstrucción completa sin depender del proveedor del modelo.

Cualquier decisión pasada se recalcula idéntica

Replay

La decisión se re-ejecuta desde su snapshot: mismo resultado, bit a bit, hoy o en cinco años.

Config drift

Si el pack cambia — un coste, un umbral, una norma — los replays del pasado lo marcan al instante. La desviación se detecta, no se descubre.

Forense

Incidentes de agentes se han descubierto meses después revisando transcripts a mano (State of AI 2026). Aquí la trazada existe por construcción.

Aprendizaje con los libros abiertos. Cada outcome recalibra las probabilidades del pack: «P pasó de 0,45 a 0,52 tras N observaciones». Cada parámetro es una fila visible y versionada — sin pesos ocultos — y los cambios de política se propagan bajo veto humano.

Historia que no se reescribe en silencio. Cuando cambie una tarifa, una ley o el comportamiento de la cartera, el replay señala exactamente qué decisiones del pasado se tomaron con los números anteriores — la trazabilidad que la supervisión pedirá en dic-2027, disponible desde el día 1.

GOBIERNO

Humano al mando — la escalera de confianza

Percibe
Consulta el expediente con permisos acotados; analiza documento y plazos (LLM confinado a lenguaje).

Propone
El motor calcula el valor de cada vía y eleva una propuesta con toda su evidencia congelada.

Firma
Una persona aprueba o descarta, con identidad y motivo. Las acciones caras siempre pasan por aquí.

Ejecuta
Solo herramientas permitidas, en sandbox aislado, con presupuesto de pasos y salida controlada.

Observa
El resultado real vuelve al motor y recalibra el pack — bajo veto humano.

Toda decisión nace PROPOSED. No existe modo automático. La autonomía se concede acción a acción, con trazada de quién la otorgó. La máquina produce; responder es siempre humano.

El agente se identifica como IA en cada interacción con clientes — AI Act, art. 50, ya vigente. Cumplimiento de fábrica, no de parche.

El AI100 de Stanford recoge un resultado incómodo: delegar decisiones en agentes sin control redujo la conducta honesta del 95 % al 12 % en experimentos. **Por eso el control aquí es arquitectura, no política.**

CATÁLOGO

Packs de decisión para banca

Pack	La decisión repetida	Coste humano hoy	Qué entrega Credence
Recobro de impagados PILOTO INMEDIATO	Miles de expedientes al día: contactar, fraccionar, escalar, derivar o esperar	4–6 € por gestión de agente; decisiones por cola, no por expediente	Estrategia por expediente con EV en euros, evidencia congelada y calibración continua con outcomes de la cartera
Reclamaciones y consultas PILOTO INMEDIATO	Plazos máximos legales, volumen alto, servicio de atención al cliente	Gestión manual durante días por expediente	Análisis del expediente y señales en minutos; borrador listo; priorización con costes de plazo; firma del gestor
Triage de alertas AML FASE 2 · CON VALIDACIÓN	Mayoría de alertas que cierran sin incidencia	Analista por alerta, colas crecientes	Cola priorizada por valor de información; cada cierre, trazado y recalibrado
Adjudicación de crédito FASE 2 · ALTO RIESGO	Evaluación de solvencia — alto riesgo Anexo III del AI Act	Comité + scoring opaco de proveedor	Capa de decisión explicable y auditable sobre el modelo existente: señales, pesos y disentimientos visibles
Fraude en pagos FASE 2	Bloquear, friccionar o pasar en tiempo real	Reglas estáticas + revisión manual	Valoración de acciones con coste de fricción del cliente en euros, decision replay para cada disputa

Un pack se siembra en semanas y se ajusta desde el admin web, sin despliegues. Cada caso maduro sube piezas genéricas al generador: el segundo vertical cuesta una fracción del primero.

INTEGRACIÓN

Agent-first: su plataforma de agentes, con el volante bien agarrado

- **MCP nativo + API/SDK.** Credence se expone como servidor MCP — el estándar de facto para herramientas de agentes en 2025-26. Claude Code, Codex, Cursor y las plataformas internas del banco lo llaman como una herramienta más.
- **Contrato de papeles nítido:** el agente percibe, redacta y explica (LLM confinado a lenguaje) · Credence calcula, registra y explica el cálculo · **la persona firma.**
- **Cuatro herramientas, nada más:** decide() · explain() · verify(record) · approve(actor)
- Toda acción del agente con consecuencias nace como **propuesta HITL** — el patrón que ya domina la producción agéntica (Decagon, PolyAI; State of AI 2026).

VISTA DEL AGENTE

```
> credence.decide(expediente=#4821)
  acción: FRACCIONAMIENTO_3P  EV 1.319 €
  probabilidades: origen declarado ✓
  estado: PROPOSED — requiere firma
> credence.approve(actor="J.R.", motivo="ok")
  registro #2.418 firmado · replay ✓
```

El LLM nunca está en el camino de decisión. Puede proponer señales nuevas y redactar; el valor de cada vía lo fija el cálculo; la ejecución exige firma. Así puede desplegar agentes con consecuencias **sin ceder el control.**

DESPLIEGUE

Instancia soberana: los datos no salen del banco

On-prem o cloud del banco
Despliegue en contenedores sobre su infraestructura; operativo incluso en red aislada (air-gap).

Sin proveedor en la decisión
Milisegundos por decisión, sin tokens en el camino de decisión. El modelo de lenguaje vive tras un gateway local — o no vive.

Stack 100 % open source
PostgreSQL, FastAPI, Keycloak, Traefik. Sin licencias runtime, sin llamadas a casa, sin dependencias del proveedor en producción.

El código del despliegue es del banco
Stack generado, estándar y legible. El expediente vive en SU PostgreSQL — sin lock-in de datos.

Contención de agentes
1 agente = 1 sandbox aislado; salida a red denegada por defecto con allowlist; presupuesto de pasos y kill switch.

Roles y seguridad a nivel de fila
Operador / aprobador / auditor con permisos diferenciados; cada escritura con actor identificado.

«Su infraestructura. Sus datos. Sus decisiones — **con nombre y apellidos.**»

La instancia corre aislada del proveedor: los upgrades llegan como artefactos firmados y es el banco quien decide cuándo aplicarlos.

EL VIENTO REGULATORIO

El AI Act llega con fecha — y el expediente no se improvisa

Ene 2025 · vigente

DORA

Gestión del riesgo TIC y de terceros para entidades financieras.

Ago 2026 · vigente

AI Act — transparencia

Aplicación general y art. 50: la IA se identifica como IA ante las personas.

Dic 2027

AI Act — Anexo III

Alto riesgo, incluida la evaluación de solvencia: explicabilidad, registro, supervisión humana (tras Reg. UE 2026/1744).

Ago 2028

AI Act — Anexo I

IA integrada en productos financieros regulados.

El punto crítico: el expediente de conformidad de dic-2027 se construye con decisiones registradas desde 2026. Quien empiece a acumular evidencia el día que la norma se exija, llegará tarde por definición.

Además, ya aplican: GDPR art. 22 (decisiones automatizadas con efectos jurídicos — derecho a explicación e intervención humana) · Directrices EBA de concesión (GL/2020/06): modelos explicables y documentados en el proceso de adjudicación · DORA para todo el ciclo TIC. **Credence produce la evidencia que estas normas presuponen.**

PROPUESTA DE ADOPCIÓN

Piloto de pago — 8 semanas, un caso, criterios congelados

S0–2

Congelar el caso

Una decisión (recobro o reclamaciones), un equipo, criterios de éxito por escrito y umbrales de NO-GO.

S2–6

Sembrar el pack

Con la cartera del banco: acciones con coste real, probabilidades con origen, validación en compilación.

S6–8

Piloto con firma

Decisiones reales PROPOSED con aprobación humana; replay sobre histórico; medición de valor por decisión.

Salida

Decidir con números

Informe de replay y drift, €/decisión vs. coste humano actual, plan de expansión — o cierre limpio.

La prueba se factura. Con criterios de éxito congelados y NO-GO automático el riesgo es compartido: nadie entra en un «piloto eterno». El coste del piloto se descuenta de la suscripción anual si hay continuidad.

Modelo comercial: suscripción por instancia + €/decisión por tramos (propuesta, ejecución, replay contados por el propio motor). El coste marginal de una decisión nueva tiende a cero; el valor por decisión, no.

La ventana está abierta

01 · FIABILIDAD

La fiabilidad es el cuello de botella nº 1 del despliegue de IA en 2026: confinar la decisión al cálculo es la vía corta a producción. Y el texto es tan frágil que su conducta cambia hasta con la identidad del personaje — **la aritmética no tiene personajes.**

02 · REGULACIÓN

Dic-2027 exige decisiones explicables, registradas y con mando humano en crédito. Los registros se acumulan desde hoy — o no existen cuando se pidan.

03 · EL FLANCO LIBRE

Los grandes proveedores no ofrecen lo que un banco necesita: responsabilidad con nombre por decisión y datos que no salen. Ese flanco está libre — hoy.

«Cuando cualquiera genera output competente en minutos, lo escaso es **quién responde por cada decisión».**

Próximo paso: una sesión de 90 minutos para congelar el caso del piloto y sus criterios de éxito. · phernandez@airflows.com · https://credence.airflows.com