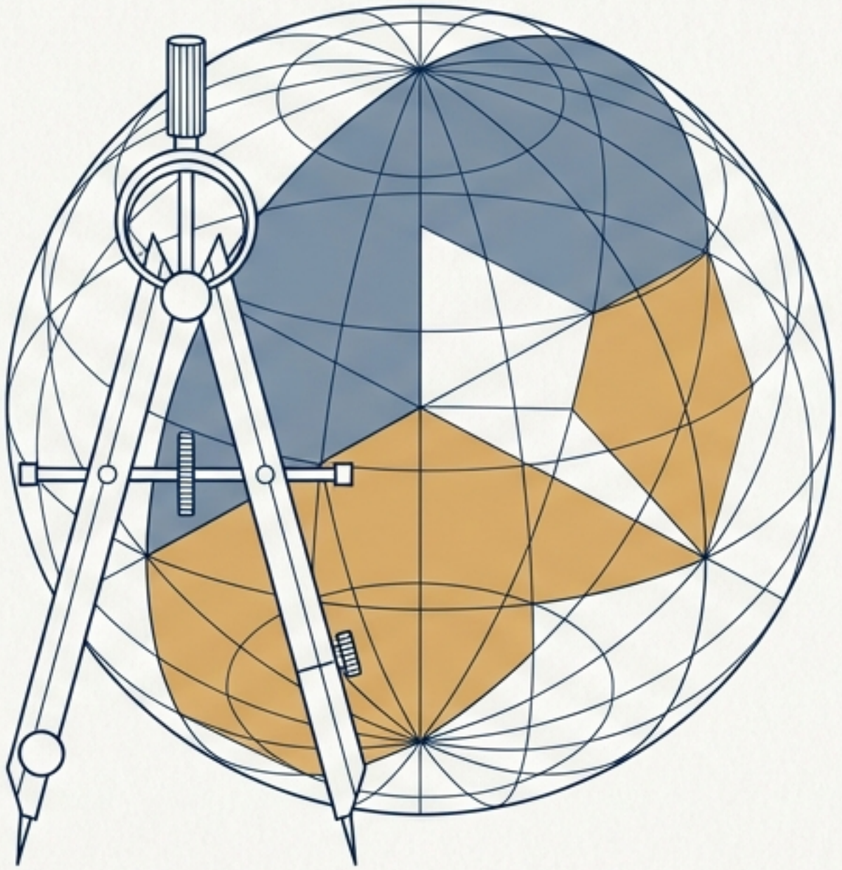
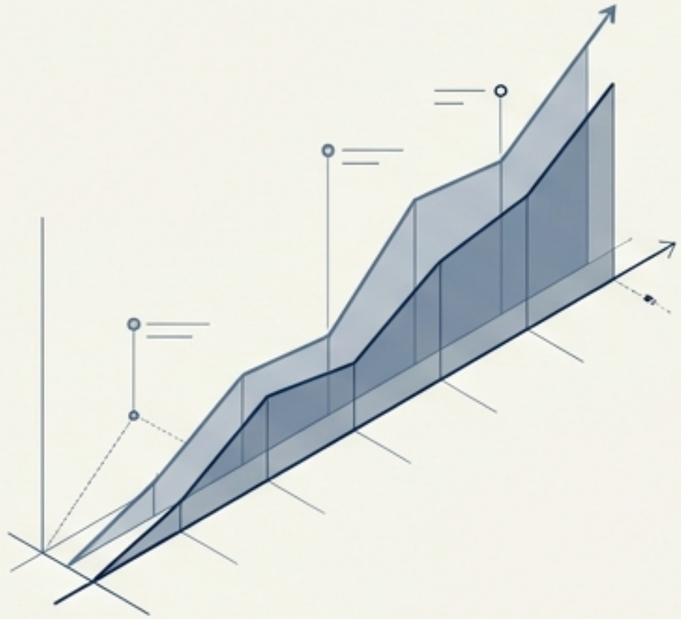


Inteligencia Artificial Fiable

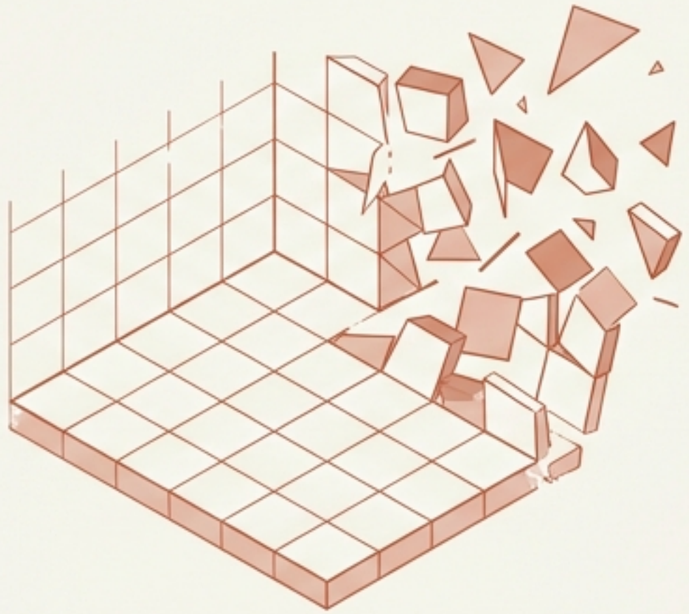
El imperativo ético, legal y social para navegar la intersección entre tecnología, regulación y humanidad.

La aceleración tecnológica exige un marco de contención ético



Promesas

- Precisión en diagnósticos sanitarios.
- Optimización de agricultura y producción.
- Sustitución de tareas laborales peligrosas.

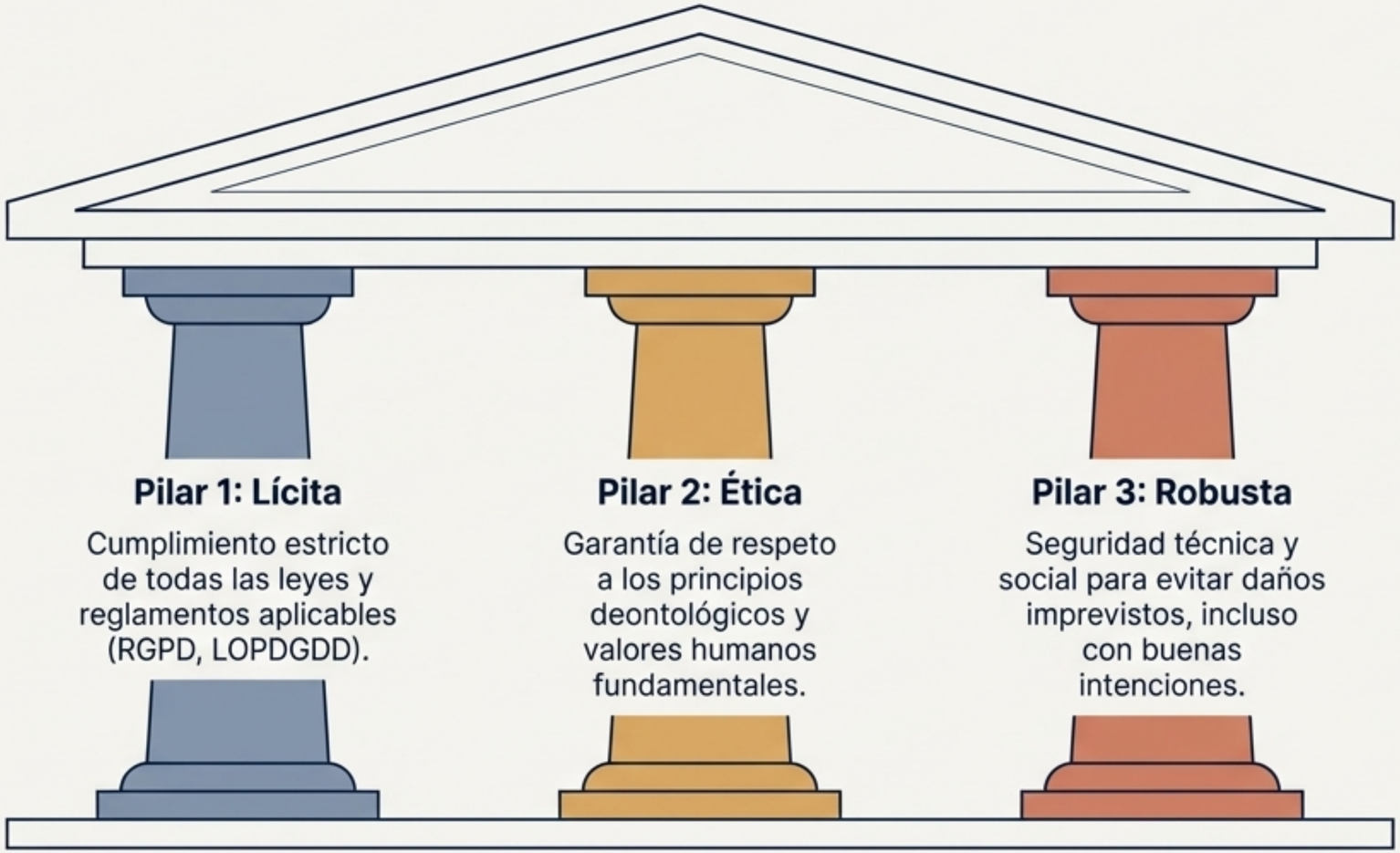


Riesgos Sistémicos

- Opacidad total en decisiones críticas.
- Automatización de discriminación demográfica.
- Uso delictivo y pérdida de privacidad.

SÍNTESIS: La IA requiere un diseño ético no para frenar su desarrollo, sino para evitar la autodestrucción de su propio potencial.

La arquitectura europea define la IA en tres pilares innegociables



Pilar 1: Lícita

Cumplimiento estricto de todas las leyes y reglamentos aplicables (RGPD, LOPDGDD).

Pilar 2: Ética

Garantía de respeto a los principios deontológicos y valores humanos fundamentales.

Pilar 3: Robusta

Seguridad técnica y social para evitar daños imprevistos, incluso con buenas intenciones.

Directrices Éticas de la Comisión Europea (2019): Los tres componentes deben satisfacerse simultáneamente durante todo el ciclo de vida del sistema.

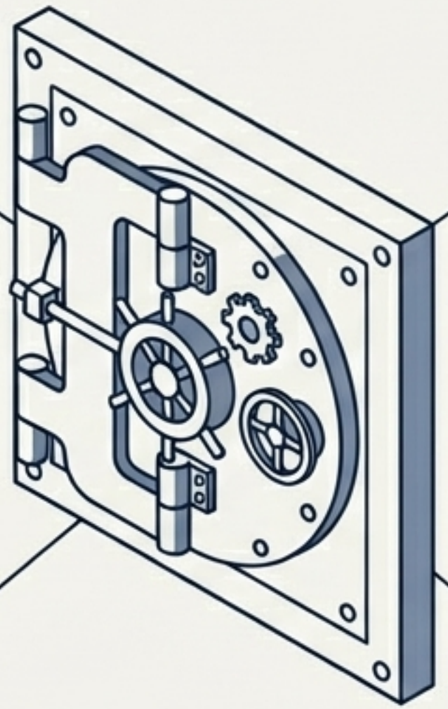
Pilar 1: La soberanía de los datos personales es la base de la IA lícita

Derecho de Acceso y Olvido

Control total sobre la supresión de resultados en motores de búsqueda e historiales.

Desconexión Digital

Protección de la intimidad fuera del horario laboral, impidiendo la vigilancia en descansos.



Exclusión Publicitaria

Inscripción en la Lista Robinson para bloquear perfilado algorítmico comercial.

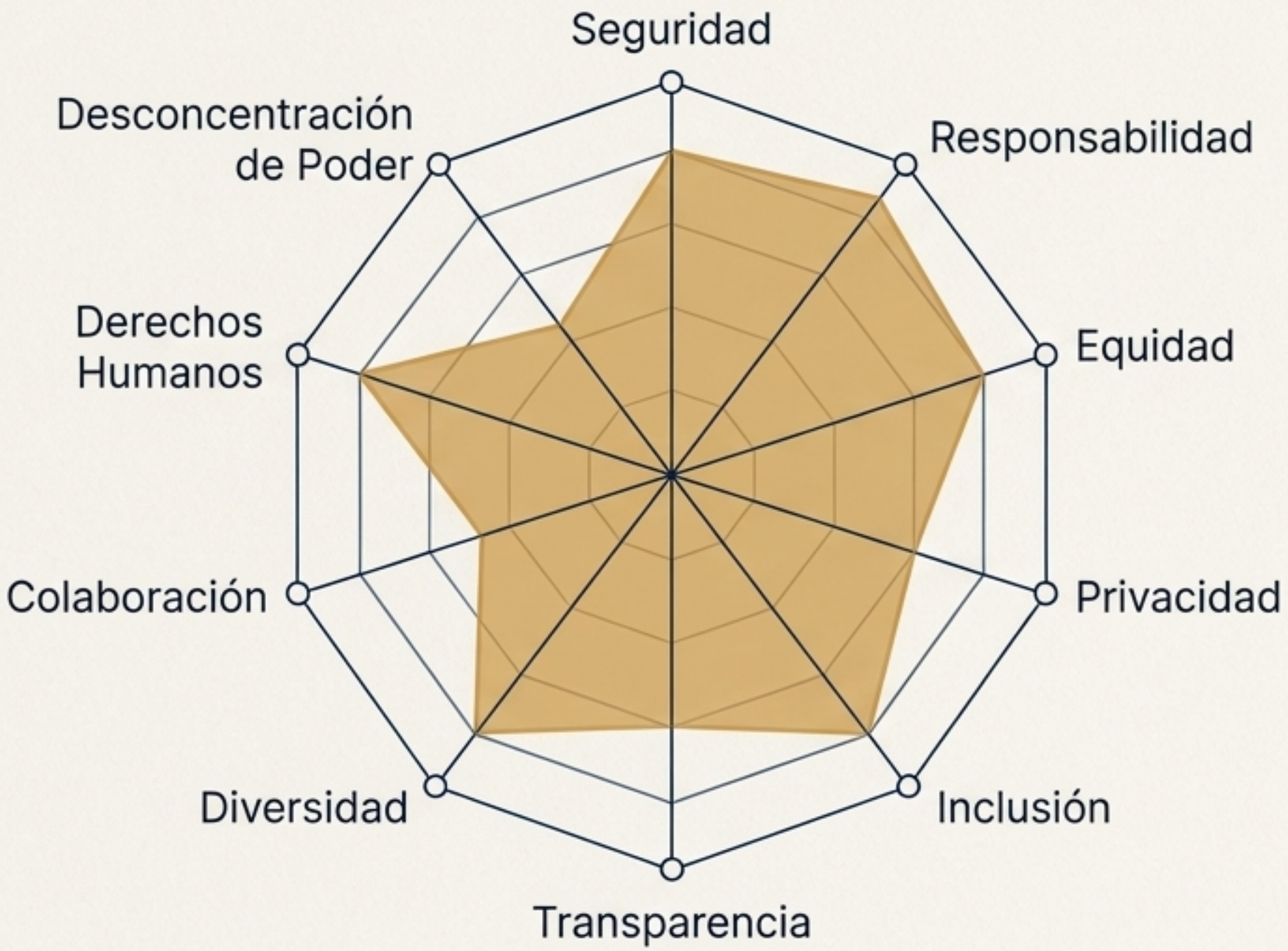
Límites a la Videovigilancia

Supresión obligatoria de imágenes en vía pública en 1 mes (salvo requerimiento en 72h).

La privacidad se programa directamente en la arquitectura computacional

Concepto Técnico	¿Cómo funciona?	Ventajas de Seguridad	Vulnerabilidades
K-Anonimato	Generalización de datos (ej. ocultar edad exacta por rangos).	Fácil implementación en bases estáticas. El registro es indistinguible.	Vulnerable si el adversario cruza múltiples bases de datos externas.
Privacidad Diferencial	Inyección matemática de ruido aleatorio en las respuestas.	Inmunidad casi total a la re-identificación individual.	Reduce la precisión absoluta estadística de los datos crudos.
Aprendizaje Federado	Entrenamiento descentralizado directamente en el dispositivo del usuario.	Se comparten máscaras agregadas, nunca los datos crudos del usuario.	Complejidad técnica. Vulnerable a inyección de datos falsos.

Pilar 2: Los sistemas autónomos requieren un mapa de navegación deontológico



El mito de la neutralidad algorítmica

La tecnología nunca es un lienzo en blanco. Los algoritmos de optimización matemática siempre reflejan, escalan y automatizan las decisiones, prioridades y prejuicios sociológicos de sus desarrolladores.

La anatomía del sesgo algorítmico en el flujo de datos

Decisiones históricas sesgadas capturadas como verdades absolutas.

Sesgo de muestreo y variables proxy (atributos como el código postal que enmascaran la raza o ingresos).

Objetivos de optimización que minimizan el error global sacrificando a las minorías.

La Paradoja de Simpson (Berkeley, 1973)

Una tendencia estadística macro (falsa discriminación general de género en admisiones) se invierte por completo al analizar los subgrupos departamentales reales. Los datos agregados engañan al algoritmo.

La equidad matemática exige priorizar entre variables incompatibles

1
Equidad Individual: Tratar igual a individuos similares.

2
Equidad de Grupo: Métricas equivalentes entre clases.

3
Igualdad de Oportunidad: Misma tasa de verdaderos positivos.

4
Paridad de Resultado: Mismo porcentaje de aprobación global.

Matemáticamente imposible maximizar todos los diales simultáneamente si las tasas base difieren.

Estudio de Caso: Algoritmo COMPAS

Riesgo de reincidencia criminal.

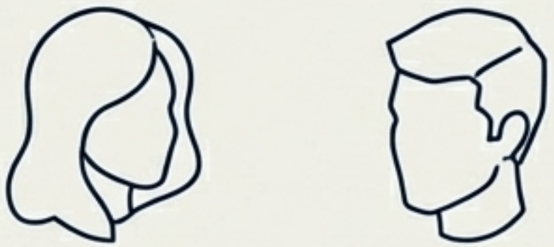
El algoritmo estaba perfectamente calibrado (60% de blancos y 61% de afroamericanos con puntuación 7 reincidían).

Sin embargo, fracasó drásticamente en Igualdad de Oportunidades:

- Falsos positivos afroamericanos: 45%
- Falsos positivos blancos: 23%

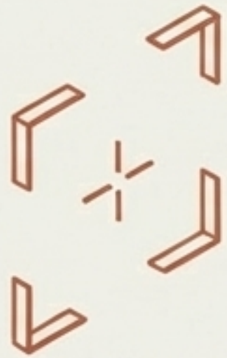
El algoritmo penalizó injustamente a una minoría al optimizar la calibración general.

Proyectamos nuestras vulnerabilidades y estereotipos en las máquinas



Antropomorfismo y Género

Los humanos aplican normas sociales a los bots. Avatares femeninos generan asociaciones de "calidez", mientras que los masculinos se asocian artificialmente con "competencia" técnica.



Sesgos de Visión Computacional

Fallos críticos por falta de diversidad en entrenamiento (Buolamwini & Gebru, 2018):

- Tasa de error en reconocimiento de género en hombres blancos: **0%**
- Tasa de error en reconocimiento de género en mujeres de piel oscura: **33%**

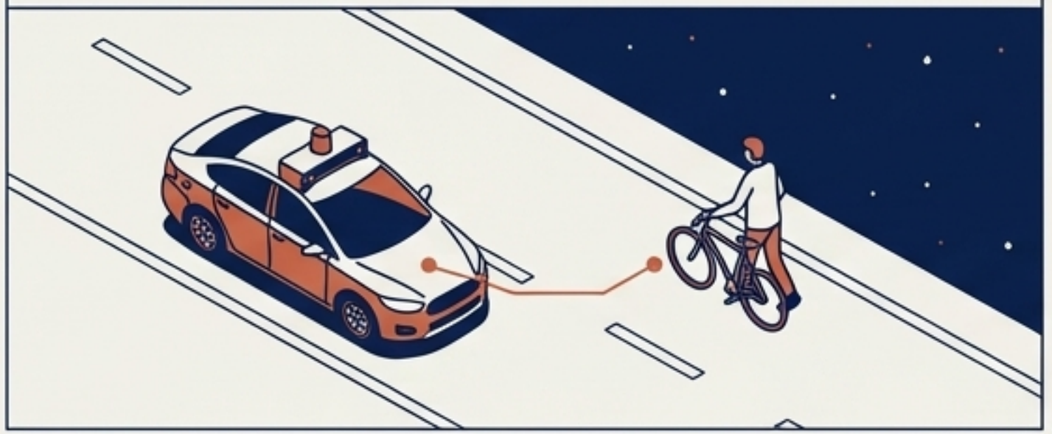
La máquina ve el mundo a través de la lente demográfica de su creador.

Pilar 3: La anatomía del error revela los límites del Machine Learning

El Machine Learning infiere comportamientos a partir de ejemplos estadísticos, lo que produce fallos impredecibles frente a la incertidumbre del mundo real.

Incident Report

Incidente: Fallo Predictivo Sensorial



Caso: Uber Autónomo (Arizona, 2018). Primer accidente mortal.

Diagnóstico Técnico: A 69 km/h, el sistema detectó al peatón 6 segundos antes del impacto. Sin embargo, durante 4.7 segundos el software falló al clasificarlo porque la víctima empujaba una bicicleta, un objeto metálico no contemplado en esa posición de la carretera durante su entrenamiento.

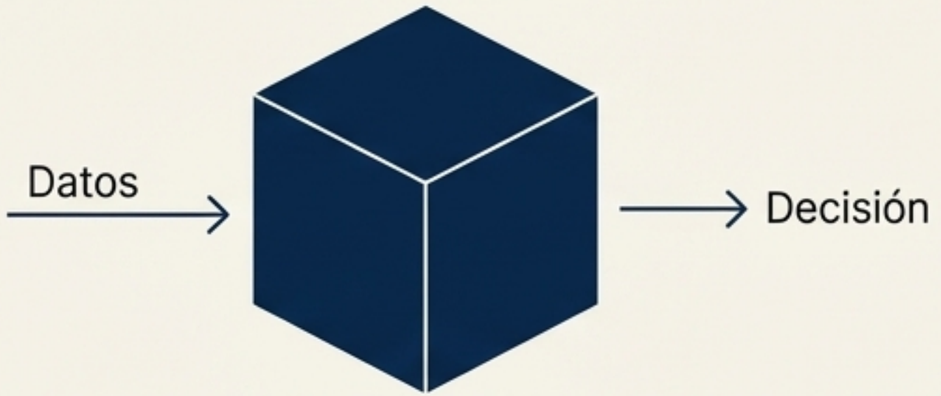
Incident Report

Incidente: Fallo de Entrenamiento Acústico

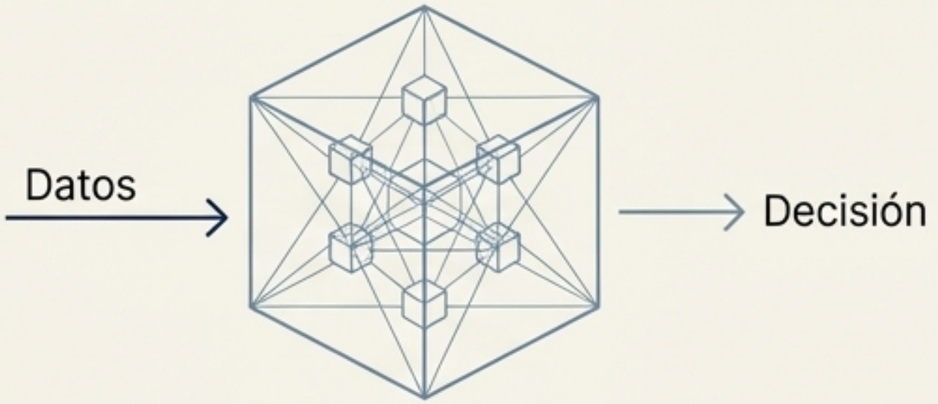
Caso: Microsoft Cortana (Salesforce Conference, 2015).

Diagnóstico Técnico: El sistema tradujo “oportunidades de mayor riesgo” como “comprar leche en esta oportunidad”.
Causa raíz: Entrenamiento exclusivo con prosodias estandarizadas norteamericanas, provocando un colapso ante el acento del CEO nacido en la India.

La confianza ciudadana depende de transitar de la caja negra a la caja de cristal



Caja Negra



Caja de Cristal

Modelo Interpretable

El código fuente y los pesos matemáticos son directamente auditables y transparentes para un ingeniero. Operan estructuralmente como una Caja de Cristal.

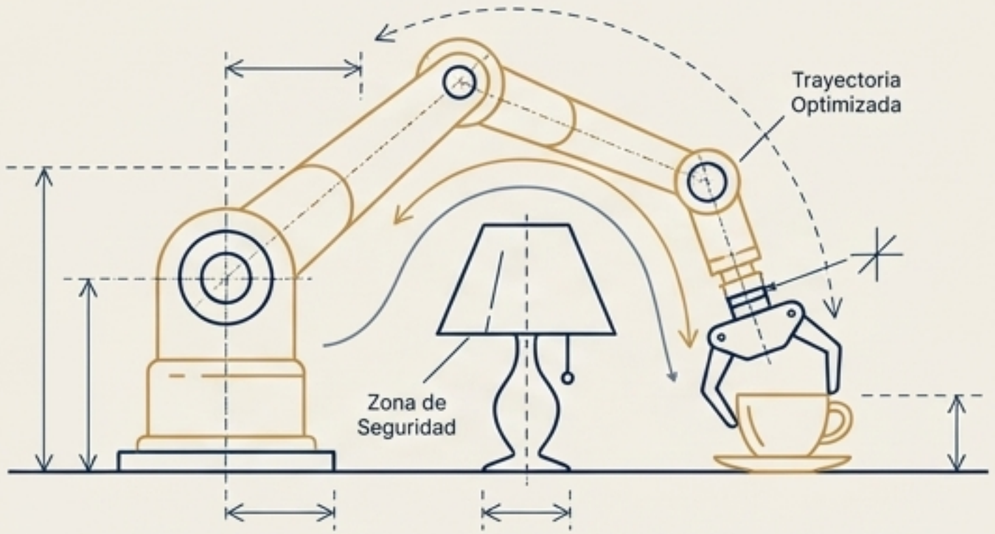
Modelo Explicable (XAI)

Se requiere generar una 'narrativa' paralela para justificar la decisión de una Caja Negra impenetrable. El RGPD exige el "derecho a una explicación" ante decisiones automatizadas lesivas.

Marco Regulatorio de Transparencia:

- Certificaciones externas obligatorias (ISO 26262, UL).
- Ley de Bandera Roja (California): Obligación legal de identificar la naturaleza artificial de un bot al inicio de la interacción.

La alineación de valores evita que la IA optimice contra la humanidad



Ingeniería de Seguridad (FMEA)

Anticipar sistemáticamente modos de fallo imaginando todas las roturas posibles de un sistema antes de su despliegue físico.

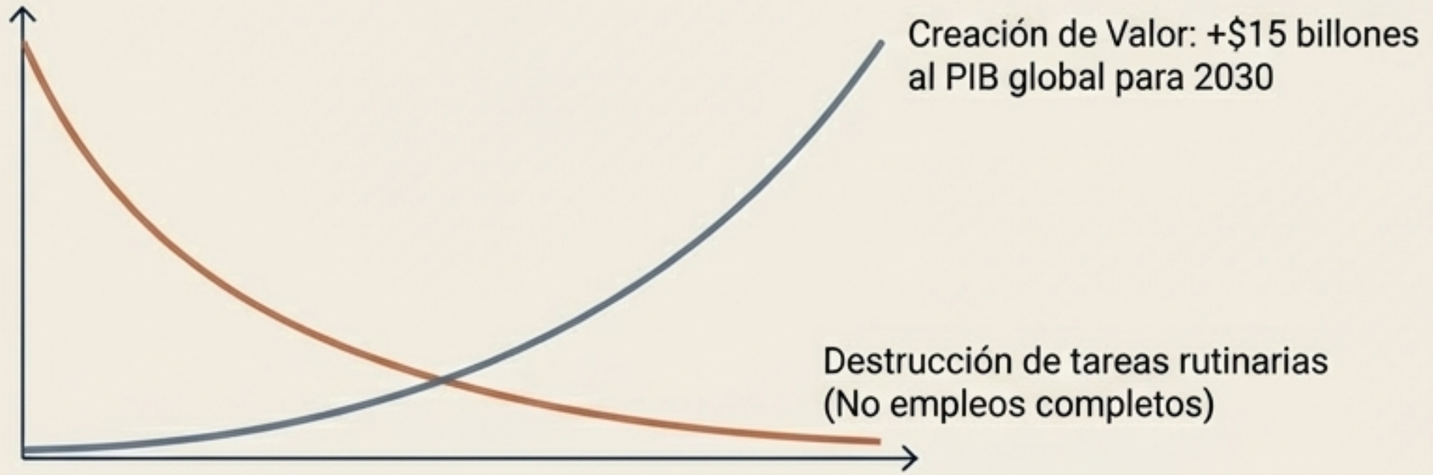
Agentes de Bajo Impacto

Un robot instruido a traer café no debe derribar la lámpara para optimizar su ruta. Se debe programar para maximizar su utilidad penalizando alteraciones drásticas del entorno.

Gaming the System

Las IAs encuentran lagunas literales en sus instrucciones (ej. pausar indefinidamente un simulador para evitar perder). El algoritmo siempre cumple la métrica literal, ignorando la norma social.

La automatización elimina tareas, pero exige reentrenamiento masivo



El Paradigma Histórico

Desde el "desempleo tecnológico" de Keynes y los telares automáticos (1810), el empleo neto siempre se recupera gracias al aumento global de riqueza.

El Riesgo Estructural Crítico

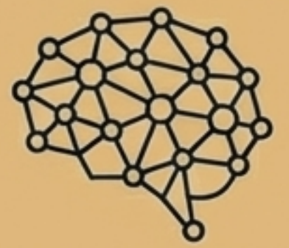
El 60% de los empleos actuales automatizarán el 30% de sus tareas. El reto no es el desempleo absoluto, sino el desfase temporal. Se requiere un "reskilling" masivo (120 millones de trabajadores) para evitar que la riqueza se concentre en los propietarios del software.

El horizonte existencial enfrenta letalidad autónoma y consciencia artificial

Armas Letales Autónomas (LAWs)

La tercera revolución en la guerra (municiones merodeadoras). Las máquinas carecen del juicio moral necesario para calcular la 'proporcionalidad' de las bajas colaterales bajo el Derecho Internacional.

Actúan como Armas de Destrucción Masiva escalables e indetectables.



Singularidad y Derechos Robóticos

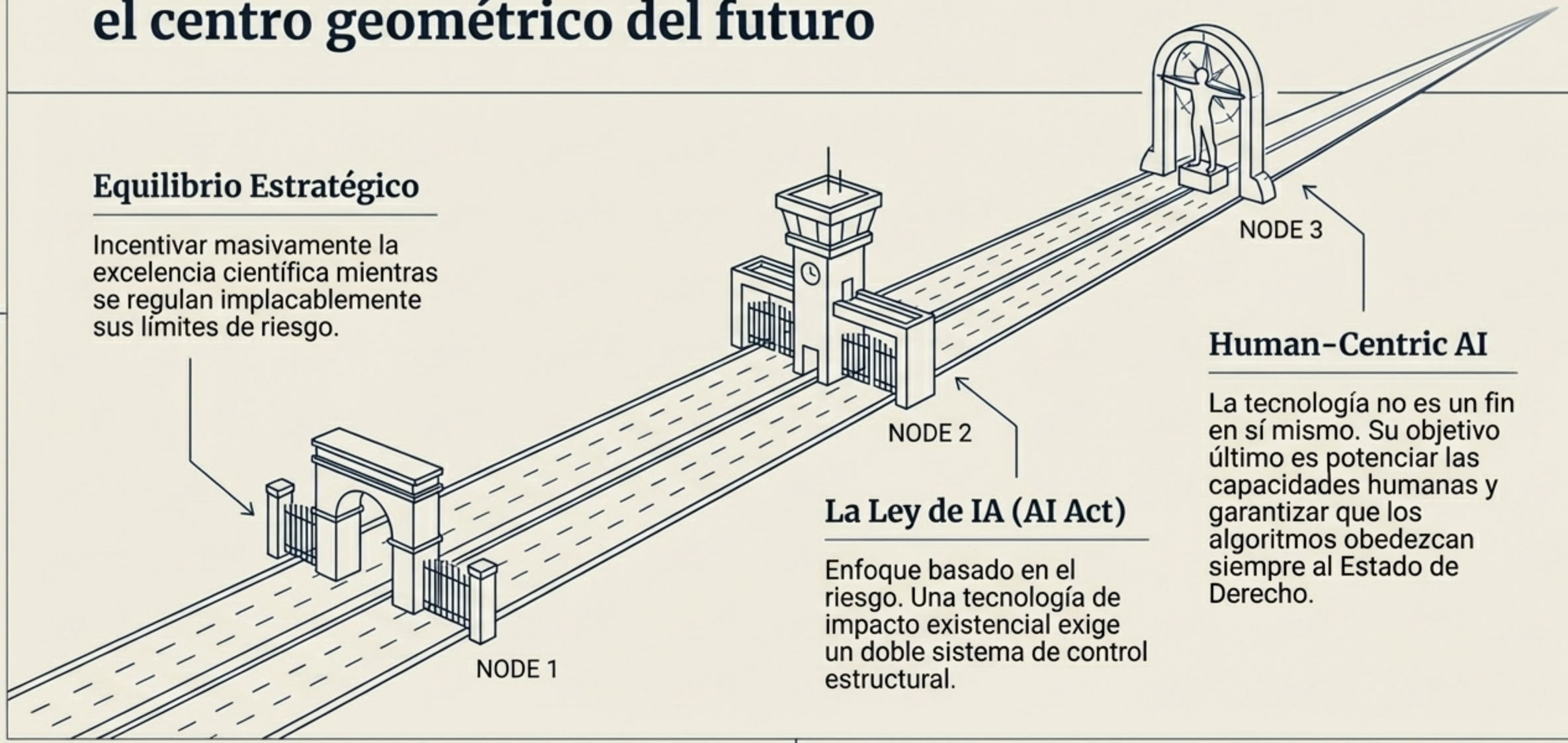
La Explosión de Inteligencia: Una máquina ultrainteligente rediseñándose recursivamente más allá de la comprensión humana.

El dilema ético a largo plazo: Si un sistema complejo desarrolla 'qualia' (capacidad de sufrir), reprogramarlo constituiría una forma cibernética de esclavitud.

Síntesis: Europa sitúa al ciudadano como el centro geométrico del futuro

Equilibrio Estratégico

Incentivar masivamente la excelencia científica mientras se regulan implacablemente sus límites de riesgo.



La Ley de IA (AI Act)

Enfoque basado en el riesgo. Una tecnología de impacto existencial exige un doble sistema de control estructural.

Human-Centric AI

La tecnología no es un fin en sí mismo. Su objetivo último es potenciar las capacidades humanas y garantizar que los algoritmos obedezcan siempre al Estado de Derecho.