

INTELIGENCIA ARTIFICIAL CON ALMA: MODELO HUMANISTA PARA UNA TRANSFORMACIÓN DIGITAL ÉTICA, CONFIABLE Y ORIENTADA AL BIENESTAR

*Artificial Intelligence with Soul: A Humanistic Model for Ethical, Trustworthy
and Well-Being-Oriented Digital Transformation*

Recibido: 31-07-2026

Aceptado: 29-08-2026

Publicado: 25-09-2026

DOI: <https://doi.org/10.53485/ret.v5i3.795>

Ignacio Bonasa Alzuria

LIDERARTE Factoría de Talento, España.

ibonasa@liderarte.org

 <https://orcid.org/0009-0001-3940-4278>

Resumen

Objetivo: formular y fundamentar el modelo «Inteligencia Artificial con Alma» como marco humanista de gobernanza, diseño, adopción y evaluación de sistemas de inteligencia artificial en organizaciones y universidades. **Metodología:** se realizó una investigación teórico-propositiva apoyada en una revisión integradora de literatura sobre inteligencia artificial confiable, ética algorítmica, diseño centrado en el ser humano, sistemas sociotécnicos, bienestar digital, justicia de datos, explicabilidad, supervisión humana y gestión de riesgos. Se compararon marcos regulatorios y técnicos internacionales, entre ellos la Recomendación de UNESCO sobre ética de la IA, el NIST AI Risk Management Framework y el Reglamento Europeo de Inteligencia Artificial. **Resultados:** se propone un modelo multinivel compuesto por seis principios —propósito legítimo, dignidad y derechos, agencia humana, justicia e inclusión, confiabilidad sociotécnica y bienestar sostenible—, cuatro capas de gobierno —estratégica, organizacional, técnica y relacional— y un ciclo de activación 4A: Aprendizaje, Actitud, Alma y Acción. El artículo aporta una matriz de evaluación, doce proposiciones contrastables, un cuadro de indicadores, una escala preliminar de 36 ítems y una hoja de ruta de implantación. **Conclusiones:** la IA responsable no puede reducirse al cumplimiento normativo ni a la precisión técnica. Requiere alinear finalidad, datos, arquitectura, decisiones, relaciones de poder y consecuencias humanas. El modelo permite traducir principios éticos en prácticas verificables y ofrece una agenda de investigación para validar su utilidad en distintos sectores y contextos culturales.

Palabras clave: inteligencia artificial humanista; ética algorítmica; gobernanza de IA; bienestar digital; transformación digital.

Abstract

Objective: To formulate and substantiate the “Artificial Intelligence with Soul” model as a humanistic framework for governing, designing, adopting, and evaluating artificial intelligence systems in organizations and universities. **Methodology:** A theoretical-propositional study was conducted through an integrative review of literature on trustworthy artificial intelligence, algorithmic ethics, human-centered design, sociotechnical systems, digital well-being, data justice, explainability, human oversight, and risk management. International regulatory and technical frameworks were compared, including UNESCO’s Recommendation on the Ethics of Artificial Intelligence, the NIST AI Risk Management Framework, and the European Union Artificial Intelligence Act. **Results:** A multilevel model is proposed comprising six principles—legitimate purpose, dignity and rights, human agency, justice and inclusion, sociotechnical trustworthiness, and sustainable well-being—four governance layers—strategic, organizational, technical, and relational—and a 4A activation cycle: Learning, Attitude, Soul, and Action. The article contributes an assessment matrix, twelve testable propositions, an indicator dashboard, a preliminary 36-item scale, and an implementation roadmap. **Conclusions:** Responsible AI cannot be reduced to regulatory compliance or technical accuracy. It requires alignment among purpose, data, architecture, decisions, power relations, and human consequences. The model translates ethical principles into verifiable practices and offers a research agenda for validating its usefulness across sectors and cultural contexts.

Keywords: humanistic artificial intelligence; algorithmic ethics; AI governance; digital well-being; digital transformation.

Introducción

La inteligencia artificial ha dejado de ser una tecnología periférica para convertirse en una infraestructura de decisión, producción de conocimiento y organización del trabajo. Sistemas de recomendación, modelos generativos, algoritmos de selección, asistentes conversacionales, analítica predictiva y automatización inteligente intervienen ya en procesos de contratación, educación, crédito, salud, comunicación, seguridad y gestión pública. Su expansión promete productividad, personalización, acceso y creatividad, pero también redistribuye poder, introduce nuevas dependencias y puede amplificar errores a una escala inédita. (Crawford, 2021; Kellogg et al., 2020; Zuboff, 2019).

El debate público suele oscilar entre dos relatos incompletos. El primero presenta la IA como solución casi automática para problemas complejos; el segundo la describe como amenaza inevitable. Ambos reducen la capacidad de gobernar. La IA no actúa en el vacío: sus efectos dependen de los fines para los que se utiliza, de los datos que la alimentan, de la arquitectura que la configura, de las instituciones que la despliegan y de las personas que reciben sus consecuencias. Por ello, la pregunta central no es si la IA es buena o mala, sino qué combinación de diseño, poder,

incentivos y supervisión produce beneficios legítimos y evita daños. (Russell, 2019; Selbst et al., 2019; Suresh & Guttag, 2021).

En los últimos años se han multiplicado principios, códigos y marcos de riesgo. UNESCO sitúa los derechos humanos y la dignidad en el centro de la ética de la IA; NIST propone gobernar, mapear, medir y gestionar riesgos; la Unión Europea adopta un enfoque regulatorio basado en niveles de riesgo. Estos avances son esenciales, pero las organizaciones siguen enfrentando una brecha de traducción: saben que deben ser transparentes, justas y responsables, aunque no siempre saben cómo convertir esas aspiraciones en decisiones de compra, diseño, implantación, seguimiento y retirada. (European Commission High-Level Expert Group on AI, 2019; Jobin et al., 2019; OECD, 2019).

El problema se agrava cuando la transformación digital se mide únicamente por velocidad, reducción de costes o volumen de automatización. Una organización puede implantar IA y deteriorar simultáneamente la autonomía, el aprendizaje, la confianza y la calidad del trabajo. También puede cumplir formalmente una norma sin comprender cómo cambia la relación entre personas y sistemas. La responsabilidad exige ampliar la unidad de análisis desde el algoritmo hacia el sistema sociotécnico completo.

Este artículo propone el modelo «Inteligencia Artificial con Alma» (IAcA). El término alma se utiliza como metáfora académica y organizacional de propósito, conciencia, dignidad, coherencia y sentido. No atribuye cualidades espirituales a la máquina. Al contrario, recuerda que la tecnología carece de finalidad moral propia y que la responsabilidad permanece en las personas e instituciones que la desarrollan, adquieren y utilizan.

El objetivo general es construir un marco humanista que integre ética, gobernanza, riesgo, bienestar y ejecución. Los objetivos específicos son: identificar el vacío entre principios y práctica; comparar marcos internacionales; definir principios y capas de gobierno; proponer mecanismos de implantación e indicadores; y formular una agenda de validación empírica.

Las preguntas de investigación son: ¿qué elementos necesita un modelo de IA humanista para superar la fragmentación entre ética, tecnología y gestión?; ¿cómo pueden operacionalizarse dignidad, agencia, justicia y bienestar durante el ciclo de vida de la IA?; ¿qué mecanismos organizacionales permiten transformar principios en decisiones verificables?; y ¿cómo evaluar si la IA amplía capacidades humanas en lugar de sustituirlas o degradarlas?

Planteamiento del problema: de la automatización a la gobernanza sociotécnica

Las organizaciones suelen tratar la adopción de IA como un proyecto tecnológico. Esta categorización facilita la asignación presupuestaria, pero oculta que todo sistema algorítmico modifica roles, criterios de decisión, acceso a información y distribución de responsabilidades. Incluso cuando el modelo no toma una decisión final, puede estructurar lo que el decisor ve, el orden de las alternativas y el nivel de atención que recibe cada caso.

La primera fuente de riesgo es la finalidad ambigua. Un sistema puede optimizar una métrica sin representar adecuadamente el valor que se pretende proteger. Por ejemplo, reducir el tiempo de atención no equivale necesariamente a mejorar el servicio; predecir rotación no legitima intervenir de cualquier manera; aumentar productividad no demuestra que el trabajo sea sostenible. La optimización algorítmica convierte objetivos mal formulados en resultados consistentemente equivocados. (Floridi, 2013; Floridi & Cowls, 2019; Taddeo & Floridi, 2018).

La segunda fuente es la opacidad distribuida. La falta de explicabilidad no reside solo en modelos complejos. También aparece en contratos, cadenas de proveedores, procedencia de datos, actualizaciones, configuraciones y decisiones de integración. Una organización puede utilizar un sistema sin comprender plenamente sus límites, y un usuario puede recibir una decisión sin saber qué información influyó ni cómo reclamar. (Pasquale, 2015; Kroll et al., 2017; Mitchell et al., 2019).

La tercera es el sesgo sociotécnico. Los datos reflejan desigualdades, omisiones y categorías históricas. Sin embargo, el sesgo no se resuelve únicamente equilibrando muestras. También surge de elegir variables proxy, definir umbrales, seleccionar poblaciones objetivo y decidir qué errores son tolerables. La justicia algorítmica es, por tanto, una cuestión técnica, política y moral.

La cuarta fuente es la erosión de la agencia. La automatización puede liberar tiempo y reducir tareas repetitivas, pero también provocar dependencia, deskilling, obediencia a recomendaciones y pérdida de juicio profesional. La supervisión humana nominal no garantiza control efectivo si la persona carece de tiempo, conocimiento, autoridad o información para contradecir al sistema.

La quinta es la externalización de costes. Los beneficios pueden concentrarse en la organización mientras los riesgos recaen sobre trabajadores, clientes, comunidades o medio

ambiente. El consumo energético, la extracción de datos, el trabajo invisible de etiquetado, la vigilancia y la precarización forman parte del balance ético.

Estas fuentes muestran que la IA responsable no es una característica intrínseca del modelo. Es una propiedad emergente del sistema completo, incluida la cultura que lo rodea. El enfoque «con alma» parte de esta premisa y desplaza la pregunta desde «¿funciona el algoritmo?» hacia «¿funciona legítimamente para las personas afectadas y bajo qué condiciones?».

Fundamentos teóricos

Inteligencia Artificial Confiable y Gestión de Riesgos

La IA confiable integra validez, resiliencia, privacidad, transparencia, explicabilidad y equidad. Bajo los marcos del NIST e ISO/IEC, la gestión de riesgos no constituye un hito puntual (p. ej., una auditoría aislada), sino un proceso continuo a lo largo del ciclo de vida del sistema. La confiabilidad es eminentemente contextual: el nivel de precisión y control requerido depende de la gravedad del daño potencial, la escala de despliegue, la vulnerabilidad de los afectados y la reversibilidad de las decisiones (National Institute of Standards and Technology, 2023; ISO/IEC, 2023).

Derechos Humanos, Dignidad y Enfoque Humanista

El marco normativo global encabezado por la Recomendación de la UNESCO, establece que el desarrollo tecnológico debe centrarse en los derechos humanos. La dignidad actúa como un límite infranqueable: las personas no deben ser reducidas a variables meramente calculables ni perder su agencia. El enfoque humanista subordina la eficiencia técnica a fines legítimos, exigiendo proporcionalidad (demostrar la necesidad del sistema y la ausencia de alternativas menos invasivas) y la preservación del juicio, autonomía y bienestar humano (UNESCO, 2021; Pirson, 2017; Sen, 1999).

Regulación Basada en Riesgo

El Reglamento Europeo de Inteligencia Artificial (AI Act) articula una gobernanza legal clasificada según el nivel de riesgo del uso tecnológico (prohibidos, alto riesgo, transparencia y propósito general). Este enfoque convierte la clasificación operativa en una decisión de gobierno que exige documentación y supervisión. No obstante, la regulación establece un suelo mínimo indispensable, pero no agota la responsabilidad ética ni el impacto real sobre la organización (European Parliament and Council, 2024).

Diseños Sociotécnicos y Centrados en el Ser Humano

Desde el diseño centrado en el ser humano y la teoría sociotécnica, los sistemas de IA no pueden evaluarse aislados del contexto relacional, las jerarquías, los incentivos y las rutinas de trabajo en que se insertan. La evaluación abarca tanto a usuarios directos como a personas indirectamente afectadas o clasificadas por la herramienta. El diseño sociotécnico analiza la distribución de tareas entre persona y máquina, velando por conservar las capacidades críticas humanas y definir con claridad las responsabilidades (Shneiderman, 2022; van de Poel, 2020).

Justicia de Datos, Explicabilidad y Contestabilidad

La equidad algorítmica va más allá de la paridad estadística. Exige decisiones normativas sobre qué métrica priorizar según el daño a prevenir (Barocas et al., 2023; Buolamwini & Gebru, 2018). Por su parte, la explicabilidad debe ser funcional al destinatario: no basta con transparentar la complejidad técnica, sino facilitar a los afectados la comprensión de las variables determinantes. La contestabilidad completa este esquema otorgando poder normativo: la posibilidad real, mediante canales accesibles y revisión humana competente, de impugnar y corregir decisiones algorítmicas (Wachter et al., 2018; Venkatasubramanian & Alfano, 2020).

Bienestar Digital, Trabajo y Capacidades

El despliegue de la IA impacta directamente en la carga cognitiva, la intensidad laboral, la autonomía y el bienestar. Basado en el enfoque de capacidades, una adopción responsable busca ampliar las posibilidades valiosas de las personas (creatividad, aprendizaje y juicio) en lugar de reducir su rol a una validación mecánica o acelerar ritmos de rendimiento (Deci & Ryan, 2000; Sen, 1999; Green, 2022).

Metodología

Se adoptó un diseño de investigación teórico-propositiva con revisión integradora. Este tipo de estudio es adecuado cuando el objetivo consiste en articular conceptos dispersos, identificar vacíos y formular un modelo susceptible de validación posterior. No se pretende demostrar eficacia causal ni presentar resultados empíricos.

La revisión se estructuró en seis dominios: ética y derechos; confiabilidad y riesgo; gobernanza; diseño sociotécnico; justicia, explicabilidad y contestabilidad; y bienestar humano. Se consideraron documentos normativos y técnicos de organismos internacionales, normas y

marcos de referencia, artículos conceptuales, revisiones y estudios seminales (Mittelstadt et al., 2016; Lepri et al., 2018; Mittelstadt, 2019).

Los criterios de inclusión fueron: relevancia directa para el ciclo de vida de IA; definición explícita de principios o mecanismos; aplicabilidad organizacional; influencia académica o institucional; y complementariedad con otros enfoques. Se excluyeron textos meramente promocionales, afirmaciones sin respaldo y fuentes que confundían capacidad técnica con legitimidad ética.

Se elaboró una matriz con unidad de análisis, finalidad, principios, nivel de intervención, mecanismos, indicadores y limitaciones. La síntesis siguió una lógica abductiva: se compararon convergencias y tensiones, y se formularon categorías capaces de explicar la brecha entre principios y práctica. (Geburu et al., 2021; Sambasivan et al., 2021).

La calidad del modelo se evaluó mediante coherencia interna, suficiencia, parsimonia, diferenciación, utilidad, falsabilidad y sensibilidad ética. Para controlar el sesgo confirmatorio, se incluyeron críticas a la ética de principios, incompatibilidades entre métricas de equidad, límites de explicabilidad y riesgos de human-washing. (Citron & Pasquale, 2014; O'Neil, 2016).

El modelo se presenta como una propuesta abierta. Sus escalas, relaciones y efectos requieren validación empírica. El vínculo con la trayectoria profesional del autor se utiliza como fuente de problematización, no como evidencia.

Comparación de marcos internacionales

La Tabla 1 resume los principales marcos que informan la propuesta y muestra el espacio que pretende cubrir el modelo IAcA.

Tabla 1

Marcos internacionales

Marco	Unidad central	Fortaleza	Límite para la práctica organizacional	Aporte al IAcA
UNESCO Ética de IA	Derechos y sociedad	Base humanista y universal	Alta abstracción para decisiones cotidianas	Dignidad, proporcionalidad, sostenibilidad
NIST AI RMF	Gestión de riesgo	Funciones operativas y ciclo de vida	Voluntario y no sustituye deliberación ética	Gobernar, mapear, medir y gestionar

EU AI Act	Usos y riesgo legal	Obligaciones vinculantes y clasificación	Cumplimiento mínimo no cubre todo impacto humano	Proporcionalidad y controles por riesgo
OECD AI Principles	Política e innovación	Cooperación, robustez y responsabilidad	Escasa granularidad organizacional	Crecimiento inclusivo y accountability
ISO/IEC 42001	Sistema de gestión	Estructura auditable de gestión de IA	Puede burocratizarse sin cultura ética	Roles, procesos y mejora continua
Diseño centrado en humanos	Usuarios y experiencia	Participación y usabilidad	Puede omitir no usuarios afectados	Co-diseño y accesibilidad
Ética por diseño	Ciclo técnico	Integra valores desde el inicio	Riesgo de reducir valores a requisitos	Trazabilidad de decisiones éticas

Modelo Inteligencia Artificial con Alma (IAcA)

El modelo define la IA con Alma como un sistema de gobierno y diseño sociotécnico que utiliza inteligencia artificial para ampliar capacidades humanas y crear valor legítimo, manteniendo dignidad, agencia, justicia, confiabilidad y bienestar durante todo el ciclo de vida.

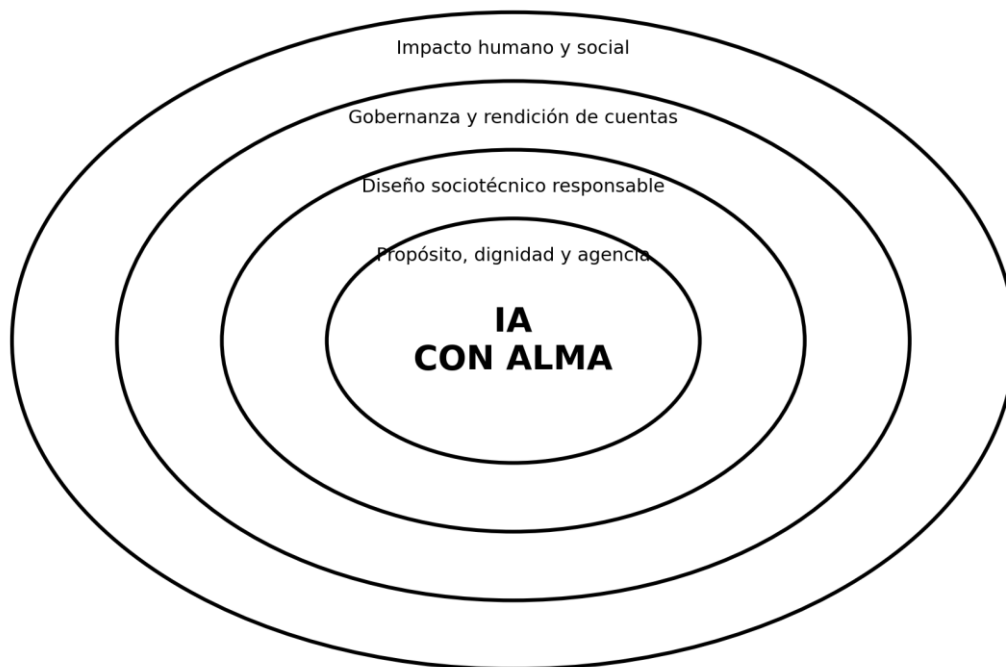


Figura 1. Arquitectura concéntrica del modelo Inteligencia Artificial con Alma.

Seis principios

Propósito legítimo. Todo sistema debe responder a una necesidad demostrable, compatible con derechos y con la misión institucional. La organización debe justificar por qué la IA es adecuada y si existe una alternativa menos invasiva.

Dignidad y derechos. Ninguna ganancia de eficiencia legítima tratar a las personas como objetos de predicción sin voz. Privacidad, no discriminación, libertad y debido proceso actúan como límites.

Agencia humana. Las personas deben conservar comprensión, capacidad de elección, competencia y autoridad. La supervisión requiere condiciones reales para intervenir y no una firma simbólica.

Justicia e inclusión. El diseño debe analizar quién se beneficia, quién soporta riesgos y quién queda fuera. Debe incluir accesibilidad, diversidad cultural y evaluación desagregada.

Confiabilidad sociotécnica. Validez, seguridad, robustez, privacidad, explicabilidad y trazabilidad se evalúan en relación con procesos, usuarios y contexto.

Bienestar sostenible. La IA debe contribuir al bienestar de personas, organizaciones, sociedad y planeta, evitando intensificación, dependencia y externalización ambiental.

Cuatro capas de gobernanza

La capa estratégica define propósito, tolerancia al riesgo, usos prohibidos, principios y rendición de cuentas. El consejo y la alta dirección deben aprobar políticas y recibir información sobre incidentes, impactos y decisiones de retirada.

La capa organizacional distribuye roles, capacidades, procesos de compra, evaluación, formación y canales de reclamación. Incluye comités multidisciplinares y participación de trabajadores y afectados.

La capa técnica integra calidad de datos, pruebas, seguridad, documentación, monitorización y control de cambios. Debe mantener trazabilidad desde la finalidad hasta el desempeño en operación.

La capa relacional analiza interacción humano-IA, confianza, carga, autonomía, accesibilidad y contestabilidad. Es donde la calidad técnica se convierte —o no— en experiencia humana legítima.

El ciclo 4A aplicado a la IA

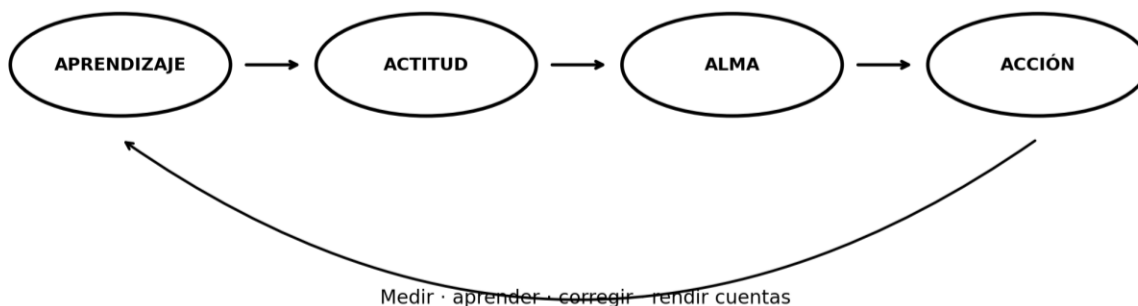


Figura 2. Ciclo 4A para la activación y mejora continua de la IA responsable.

Aprendizaje implica comprender capacidades, límites, contexto y efectos. Incluye alfabetización, evaluación de evidencia y revisión de incidentes.

Actitud representa prudencia, curiosidad, responsabilidad y disposición a cuestionar la automatización. Evita tanto el rechazo irreflexivo como la fascinación tecnológica.

Alma conecta el proyecto con propósito, dignidad, valores y sentido. Pregunta qué humanidad se desea proteger y qué relaciones se están creando.

Acción traduce principios en controles, responsabilidades, diseño, formación, auditoría, indicadores y decisiones de continuar, modificar o retirar.

Teoría del cambio

El modelo asume que los principios no producen impacto por sí mismos. Necesitan capacidades organizacionales, procesos de ciclo de vida y mecanismos de aprendizaje. La secuencia propuesta es: propósito legítimo y gobernanza; participación y diseño sociotécnico; controles técnicos y relacionales; uso acompañado; monitorización; corrección y rendición de cuentas.

Los mecanismos mediadores son comprensión, confianza calibrada, competencia humana, justicia procedimental y capacidad de contestación. Cuando están presentes, la IA puede mejorar calidad, acceso, creatividad y bienestar. Cuando faltan, la misma capacidad técnica puede generar dependencia, exclusión o daño. (Nissenbaum, 2010; Raji & Buolamwini, 2019; Raji et al., 2020).

Los moderadores incluyen criticidad, vulnerabilidad, escala, incertidumbre, reversibilidad, cultura, madurez digital y regulación. Cuanto mayor sea el riesgo, más exigentes deben ser la

evidencia, la supervisión y las vías de recurso. (National Institute of Standards and Technology, 2024).

El modelo incorpora un principio de no compensación: un alto beneficio agregado no compensa vulneraciones graves de derechos ni daños irreversibles. Este principio impide que un índice global o una mejora de productividad oculten impactos inaceptables sobre grupos específicos.

Tabla 2

Matriz de evaluación del ciclo de vida

Fase	Preguntas críticas	Evidencias mínimas	Decisión posible
Problema	¿La necesidad está demostrada? ¿Es necesaria la IA?	Caso de uso, alternativas, grupos afectados	No iniciar / redefinir
Datos	¿Son legítimos, pertinentes y representativos?	Origen, consentimiento/base legal, calidad, sesgos	Corregir / limitar
Diseño	¿Qué capacidad humana se amplía o degrada?	Requisitos, co-diseño, análisis de tareas	Rediseñar
Validación	¿Funciona para subgrupos y escenarios adversos?	Métricas, pruebas, límites, evaluación humana	Aprobar condicionado
Despliegue	¿Existen formación, supervisión y recurso?	Protocolos, responsables, información al usuario	Piloto / pausar
Operación	¿Cambian datos, conducta o contexto?	Monitorización, incidentes, feedback, deriva	Modificar / retirar
Retirada	¿Cómo se preservan derechos y continuidad?	Plan de salida, portabilidad, comunicación	Desactivar responsablemente

Proposiciones teóricas

P1. La claridad del propósito legítimo se asociará positivamente con la aceptación informada y la calidad de las decisiones de adopción.

P2. La participación temprana de personas afectadas reducirá riesgos no identificados por equipos exclusivamente técnicos.

P3. La supervisión humana efectiva mediará la relación entre explicabilidad y confianza calibrada.

P4. La contestabilidad moderará el impacto negativo de errores algorítmicos sobre confianza y legitimidad.

P5. La madurez de gobernanza fortalecerá la relación entre calidad técnica e impacto organizacional sostenible.

P6. La intensificación del trabajo mediará negativamente la relación entre automatización y bienestar.

P7. La inversión de ganancias de eficiencia en aprendizaje y autonomía aumentará el efecto positivo de la IA sobre capacidades humanas.

P8. La evaluación desagregada identificará daños ocultos que no aparecen en métricas agregadas.

P9. La coherencia entre capas estratégica, organizacional, técnica y relacional predecirá mejores resultados que el cumplimiento aislado.

P10. La alfabetización crítica reducirá la sobreconfianza y la infrautilización de sistemas de IA.

P11. La trazabilidad de decisiones éticas aumentará la responsabilidad percibida y reducirá el human-washing.

P12. El bienestar sostenible mediará parcialmente la relación entre adopción responsable y desempeño de largo plazo.

Instrumento preliminar IAcA-36

Se propone una escala de 36 ítems, seis por principio, con respuesta Likert de 1 a 7. Su finalidad es facilitar investigación futura; no debe utilizarse como instrumento validado hasta completar el protocolo psicométrico.

Tabla 3

Operacionalización IAcA

Dimensión	Ítem	Enunciado
Propósito legítimo	1	El uso de IA responde a una necesidad claramente definida.
Propósito legítimo	2	Se consideraron alternativas menos invasivas.
Propósito legítimo	3	Los objetivos del sistema son compatibles con la misión.
Propósito legítimo	4	Los indicadores representan valor real y no solo eficiencia.

Propósito legítimo	5	Se documentan límites y usos no permitidos.
Propósito legítimo	6	El sistema puede detenerse si deja de ser legítimo.
Dignidad y derechos	1	Se evalúan impactos sobre derechos fundamentales.
Dignidad y derechos	2	Las personas reciben información comprensible.
Dignidad y derechos	3	La privacidad se protege por diseño.
Dignidad y derechos	4	No se utilizan prácticas manipulativas.
Dignidad y derechos	5	Existen salvaguardas para personas vulnerables.
Dignidad y derechos	6	Los derechos actúan como límites no negociables.
Agencia humana	1	Los usuarios comprenden el papel de la IA.
Agencia humana	2	La supervisión humana puede modificar decisiones.
Agencia humana	3	Las personas conservan habilidades críticas.
Agencia humana	4	Existe tiempo para revisar recomendaciones.
Agencia humana	5	La automatización amplía autonomía.
Agencia humana	6	La responsabilidad humana está claramente asignada.
Justicia e inclusión	1	Los resultados se analizan por grupos relevantes.
Justicia e inclusión	2	Participan personas afectadas y diversas.
Justicia e inclusión	3	Se evalúa accesibilidad.
Justicia e inclusión	4	Se revisan variables proxy y sesgos.
Justicia e inclusión	5	Los beneficios y riesgos se distribuyen equitativamente.
Justicia e inclusión	6	Existen medidas correctoras ante discriminación.
Confiabilidad sociotécnica	1	El sistema ha sido validado en su contexto real.
Confiabilidad sociotécnica	2	Se monitorizan errores y deriva.
Confiabilidad sociotécnica	3	Existen controles de seguridad y privacidad.

Confiabilidad sociotécnica	4	Las explicaciones son útiles para cada destinatario.
Confiabilidad sociotécnica	5	Las decisiones y cambios son trazables.
Confiabilidad sociotécnica	6	Los proveedores ofrecen documentación suficiente.
Bienestar sostenible	1	La IA reduce carga innecesaria sin intensificar el trabajo.
Bienestar sostenible	2	Las ganancias se reinvierten en desarrollo humano.
Bienestar sostenible	3	Se protege el derecho a desconexión.
Bienestar sostenible	4	Se evalúan impactos ambientales.
Bienestar sostenible	5	La tecnología mejora la calidad de relaciones o servicios.
Bienestar sostenible	6	Se monitorizan efectos no previstos sobre bienestar.

Protocolo de validación empírica

La validación debería comenzar con entrevistas cognitivas en diversos sectores para evaluar comprensión, relevancia y ambigüedad. Un panel de expertos en IA, ética, derecho, psicometría y gestión estimaría validez de contenido mediante índices de relevancia y claridad.

Se recomienda una muestra exploratoria mínima de 300 participantes y una muestra independiente para confirmación. El análisis factorial exploratorio debería utilizar correlaciones policóricas y métodos robustos. La estructura se contrastaría mediante análisis factorial confirmatorio.

La fiabilidad se estimaría con omega y alfa, evitando depender de un único coeficiente. La validez convergente y discriminante se evaluaría mediante AVE, fiabilidad compuesta y HTMT. La estabilidad temporal requeriría test-retest.

La invarianza se examinaría por sector, género, tamaño organizacional, país y rol. Sin invarianza, las comparaciones de puntuaciones pueden ser engañosas. También conviene estudiar sesgo diferencial de ítems.

La validez de criterio se analizaría con incidentes, confianza, bienestar, calidad del servicio, rotación, reclamaciones y desempeño. Modelos de ecuaciones estructurales permitirían contrastar mecanismos y moderadores.

Los puntos de corte no deberían definirse por conveniencia. Deben derivarse de evidencia sobre riesgos y resultados. En usos de alto impacto, una puntuación total nunca debe sustituir el cumplimiento de requisitos críticos.

Hoja de ruta de implantación

1. **Mandato y límites.** Aprobar política, propósito, usos prohibidos, roles y criterios de escalado.
2. **Inventario.** Registrar sistemas, proveedores, datos, usuarios, afectados, criticidad y dependencias.
3. **Priorización.** Clasificar por riesgo legal, humano, operativo, reputacional y ambiental.
4. **Evaluación de impacto.** Analizar necesidad, derechos, equidad, bienestar, seguridad y alternativas.
5. **Co-diseño y controles.** Integrar afectados, expertos y usuarios; definir supervisión y recurso.
6. **Piloto responsable.** Desplegar a escala limitada, con métricas, seguimiento y posibilidad de reversión.
7. **Operación y monitorización.** Observar deriva, incidentes, efectos humanos y cambios de contexto.
8. **Rendición de cuentas.** Publicar información proporcional, atender reclamaciones y aprender.
9. **Revisión o retirada.** Modificar, suspender o retirar cuando el valor no justifique el riesgo.

Tabla 4

Cuadro de indicadores

Dominio	Indicadores de proceso	Indicadores de resultado	Señales de alerta
Propósito	Casos con análisis de alternativas	Valor creado para afectados	Optimización de métricas proxy
Derechos	Evaluaciones y salvaguardas	Reclamaciones resueltas	Uso desproporcionado de datos
Agencia	Formación y revisión humana	Autonomía y competencia percibida	Sesgo de automatización
Justicia	Pruebas desagregadas	Brechas de error y acceso	Daño concentrado en un grupo

Confiabilidad	Validación, monitorización, seguridad	Errores, incidentes, disponibilidad	Deriva, vulnerabilidades
Bienestar	Carga, desconexión, participación	Energía, calidad del trabajo, confianza	Intensificación y vigilancia
Sostenibilidad	Medición de recursos y proveedores	Impacto ambiental neto	Consumo creciente sin control

Aplicaciones sectoriales

Organizaciones y recursos humanos

En selección, evaluación y gestión del talento, la IA puede mejorar consistencia y reducir tareas, pero afecta oportunidades y trayectorias. El modelo exige validación relacionada con el puesto, análisis de impacto, información a candidatos, revisión humana y vías de reclamación.

En productividad, los asistentes generativos deben evaluarse por calidad, aprendizaje y carga, no solo por velocidad. La organización debe evitar convertir cada ahorro de tiempo en más volumen de trabajo.

En vigilancia laboral, la proporcionalidad es crítica. La disponibilidad técnica para medir actividad no crea legitimidad para observar de forma permanente.

Educación superior

Las universidades necesitan reglas transparentes, alfabetización para docentes y estudiantes, evaluación auténtica y protección de datos. Prohibir sin enseñar o permitir sin límites son respuestas insuficientes.

El modelo conecta con la Universidad de la Felicidad y del Bienestar al situar la tecnología como medio para el desarrollo humano y no como finalidad institucional.

Servicios públicos y sociales

En servicios públicos, la IA puede ampliar acceso y detectar necesidades, pero las personas afectadas no siempre pueden elegir otro proveedor. Por ello, la obligación de explicación, debido proceso y accesibilidad es mayor.

Los modelos no deben sustituir la responsabilidad administrativa. La contestabilidad debe incluir revisión por una persona con competencia y autoridad.

Las métricas de eficiencia deben acompañarse de equidad, confianza y calidad de atención.

Salud y bienestar (World Health Organization, 2021)

En salud, el apoyo algorítmico puede mejorar detección y personalización, pero el riesgo de daño y la sensibilidad de datos requieren evidencia clínica, validación local, supervisión y comunicación de incertidumbre.

Las aplicaciones de bienestar no deben medicalizar emociones normales ni utilizar técnicas manipulativas. La privacidad y el control del usuario son condiciones esenciales.

El modelo propone evaluar no solo precisión, sino relación terapéutica, autonomía, acceso y desigualdad.

Riesgos específicos de la inteligencia artificial generativa (UNESCO, 2023)

La inteligencia artificial generativa introduce una combinación particular de capacidad, incertidumbre y escala. A diferencia de sistemas diseñados para una tarea delimitada, los modelos de propósito general pueden producir texto, imágenes, código y recomendaciones en contextos muy distintos. Esta versatilidad amplía oportunidades, pero dificulta anticipar todos los usos y efectos. La gobernanza debe considerar tanto el modelo como la aplicación, el usuario, los datos añadidos y el entorno institucional.

La confabulación constituye uno de los riesgos más visibles. Un sistema puede producir contenido lingüísticamente convincente y factualmente incorrecto. La fluidez aumenta la probabilidad de sobreconfianza. Por ello, las organizaciones deben definir usos en los que la verificación sea viable, establecer fuentes autorizadas y comunicar incertidumbre. En decisiones de alto impacto, la salida generativa no debe convertirse en evidencia por sí misma.

La integridad de la información requiere atención especial. Los modelos pueden reproducir estereotipos, inventar referencias, sintetizar desinformación o facilitar contenidos manipulativos. La respuesta no consiste únicamente en filtros técnicos; requiere políticas de publicación, trazabilidad de fuentes, revisión editorial y alfabetización crítica.

La propiedad intelectual plantea incertidumbres sobre datos de entrenamiento, similitud de salidas, licencias y autoría. La organización debe diferenciar entre inspiración, transformación y reproducción, mantener registros de uso y evitar introducir información confidencial en herramientas no autorizadas.

La configuración humano-IA puede producir dependencia cognitiva. Cuando el sistema redacta, resume, programa o decide de manera constante, las personas pueden perder práctica y

criterio. La prevención exige diseñar tareas que mantengan comprensión, alternar asistencia y ejecución autónoma, y evaluar aprendizaje, no solo productividad.

Los riesgos de seguridad incluyen generación de código vulnerable, ingeniería social, filtración de datos y ataques de prompt. Los controles deben integrar gestión de acceso, aislamiento de información, pruebas adversariales, monitorización y planes de respuesta.

Finalmente, la sostenibilidad ambiental obliga a considerar consumo energético, infraestructura y frecuencia de uso. No toda tarea necesita el modelo más grande. La elección proporcional de recursos forma parte de la responsabilidad técnica y organizacional.

Adquisición y gobierno de proveedores

Gran parte de las organizaciones no desarrolla modelos propios, sino que adquiere servicios. Esta realidad desplaza parte del riesgo hacia contratos y dependencias. Comprar IA no transfiere la responsabilidad sobre sus consecuencias. La entidad usuaria debe conocer qué datos se procesan, dónde se alojan, cómo se actualiza el sistema y qué derechos conserva.

La diligencia debida debe comenzar antes de la selección. El proveedor debería aportar documentación sobre finalidad, datos, validación, limitaciones, seguridad, accesibilidad, subcontratistas, cambios de versión y gestión de incidentes. Las afirmaciones comerciales deben contrastarse con evidencia relevante para el contexto real.

Los contratos necesitan cláusulas de auditoría, notificación de cambios, continuidad, portabilidad, eliminación de datos, niveles de servicio y cooperación ante reclamaciones. También deben definir quién asume costes de corrección y cómo se gestiona una retirada.

El lock-in tecnológico puede convertir una decisión táctica en dependencia estratégica. La organización debe diseñar un plan de salida desde el inicio, mantener conocimiento interno y evitar que la lógica del proveedor sustituya los criterios institucionales.

La evaluación no finaliza con la firma. Los modelos y servicios cambian. Cada actualización relevante debe activar una revisión proporcional, especialmente si modifica capacidades, datos o condiciones de uso.

Auditoría algorítmica y evaluación de impacto

La auditoría algorítmica examina evidencia sobre diseño, datos, desempeño, controles y consecuencias. Puede ser interna, independiente o regulatoria. Su valor depende del alcance, acceso y capacidad de producir cambios. Una auditoría sin acceso a información relevante o sin autoridad para corregir tiene utilidad limitada.

La evaluación de impacto debe iniciarse antes del desarrollo o compra. Debe describir finalidad, afectados, datos, decisiones, beneficios, escenarios de daño, medidas y riesgo residual. No es un documento estático: se actualiza cuando cambia el sistema o el contexto.

El modelo IAcA propone combinar cinco tipos de evidencia: documentación técnica; métricas desagregadas; observación del trabajo real; experiencia de personas afectadas; e incidentes y reclamaciones. Esta triangulación evita confundir el rendimiento de laboratorio con el impacto en operación.

La independencia es gradual. Los equipos internos poseen contexto, pero pueden estar sujetos a presión. La revisión externa aporta distancia, aunque puede carecer de acceso o comprensión. Un sistema robusto combina autoevaluación, segunda línea de control y auditoría independiente para usos críticos.

Los resultados deben vincularse con decisiones. La respuesta puede incluir continuar, limitar, rediseñar, suspender o retirar. El riesgo residual debe ser aceptado por la autoridad correspondiente y no por el equipo técnico en solitario.

Tabla 4

Dimensiones de auditoría

Dimensión de auditoría	Evidencia	Pregunta de aseguramiento	Resultado
Finalidad	Caso de uso, alternativas, indicadores	¿La IA sigue siendo necesaria y proporcional?	Legitimidad
Datos	Origen, calidad, cobertura, permisos	¿Los datos son adecuados y lícitos?	Integridad
Modelo	Pruebas, límites, robustez	¿El desempeño es suficiente y estable?	Confiabilidad
Interacción	Flujos, supervisión, accesibilidad	¿La persona puede comprender e intervenir?	Agencia
Impacto	Brechas, bienestar, reclamaciones	¿Existen daños concentrados o acumulativos?	Justicia
Gobierno	Roles, incidentes, cambios	¿Hay responsabilidad y capacidad de respuesta?	Accountability

Caso ilustrativo: asistente generativo para una universidad con alma

Para ilustrar el modelo, se presenta un caso hipotético, no empírico. Una universidad desea implantar un asistente generativo para orientar a estudiantes, apoyar a docentes y reducir tiempos administrativos. El proyecto se inicia con el propósito de ampliar acceso y acompañamiento, no de sustituir tutoría ni decisión académica.

En la fase de aprendizaje, el equipo identifica usos, datos y poblaciones. Descubre que estudiantes internacionales, personas con discapacidad y quienes cursan en segunda lengua pueden beneficiarse, pero también enfrentar respuestas incorrectas o inaccesibles. Se definen usos permitidos: orientación general, búsqueda en normativa verificada, traducción preliminar y apoyo a la planificación. Se excluyen decisiones de admisión, calificación, disciplina y salud mental.

La actitud institucional se concreta en prudencia y experimentación. El sistema se pilota con un grupo limitado, se informa que es una herramienta automatizada y se ofrece contacto humano. Las respuestas incluyen fuentes y advertencias. Los datos sensibles no se utilizan para entrenamiento sin base y consentimiento adecuados.

El alma del proyecto se expresa en cuatro compromisos: ningún estudiante quedará sin alternativa humana; la herramienta no manipulará ni perfilará para fines comerciales; la diversidad lingüística será valor de diseño; y las ganancias de eficiencia se dedicarán parcialmente a tutoría y apoyo humano.

La acción incluye pruebas de precisión por idioma, accesibilidad, monitorización de preguntas no resueltas, revisión de sesgos y canales de reporte. Los responsables pueden suspender funciones. El consejo recibe indicadores de uso, error, satisfacción, equidad y carga del personal.

El caso muestra que la calidad no depende solo del modelo. Depende de fuentes, límites, interfaz, formación, rutas humanas y reinversión de beneficios. La misma tecnología, sin estos elementos, podría convertirse en barrera automatizada.

Tabla 5

Lista de comprobación para una IA con Alma

Área	Pregunta mínima antes del despliegue
Propósito	¿Se ha demostrado el problema y considerado una alternativa no algorítmica?
Derechos	¿Existe algún derecho que el sistema pueda limitar o vulnerar?
Participación	¿Han intervenido personas afectadas, incluidas las más vulnerables?

Datos	¿Se conoce la procedencia, calidad, licitud y representatividad?
Desempeño	¿Las métricas reflejan el valor real y se analizan por subgrupos?
Explicación	¿Cada destinatario recibe información comprensible y útil?
Supervisión	¿La persona tiene tiempo, formación y autoridad para contradecir?
Recurso	¿Se puede cuestionar y corregir una decisión con rapidez?
Bienestar	¿Se ha medido carga, autonomía, aprendizaje y confianza?
Proveedor	¿Hay auditoría, notificación de cambios y plan de salida?
Incidentes	¿Existe un protocolo de detección, comunicación y reparación?
Retirada	¿Se puede desactivar el sistema sin perjudicar la continuidad?

Discusión

Posicionamiento del modelo frente a la literatura sobre ética de la IA

La proliferación de declaraciones éticas sobre inteligencia artificial constituye, paradójicamente, el punto de partida del problema que este artículo aborda. El mapeo global de Jobin et al. (2019) documentó una convergencia nominal en torno a transparencia, justicia, no maleficencia, responsabilidad y privacidad, pero también una divergencia sustantiva en la interpretación de cada término y en los mecanismos previstos para su cumplimiento. Floridi y Cows (2019) intentaron reducir esa dispersión a cinco principios unificados; Mittelstadt (2019), en contraste, argumentó que la analogía con la bioética es defectuosa, porque la IA carece de fines comunes acordados, de historia profesional compartida, de métodos probados para traducir principios en casos y de mecanismos de responsabilidad legal equivalentes. Hagendorff (2020) añadió que buena parte de las guías existentes carece de cualquier dispositivo de refuerzo, por lo que funcionan como declaraciones reputacionales antes que como instrumentos de gobierno.

El modelo IAcA se sitúa deliberadamente en esa brecha. No propone un séptimo catálogo de principios, sino una arquitectura de traducción: seis principios asociados a cuatro capas de gobierno, un ciclo de activación (4A), una matriz de decisión por fase del ciclo de vida, doce proposiciones contrastables y un instrumento de medida preliminar. La diferencia respecto de la ética de principios no es el contenido normativo —que coincide en gran medida con UNESCO

(2021), OECD (2019) o el European Commission High-Level Expert Group on AI (2019), sino la exigencia de que cada principio se asocie a una evidencia mínima, un responsable identificado y una decisión posible, incluida la de no desplegar. En términos de Dignum (2019), se trata de desplazar la pregunta desde *qué valores debe respetar la IA* hacia *qué dispositivos organizacionales hacen que esos valores alteren efectivamente una decisión*.

La unidad de análisis: del algoritmo al sistema sociotécnico

La segunda contribución del modelo consiste en ampliar la unidad de análisis. Selbst et al. (2019) identificaron cinco «trampas de abstracción» en las que incurre el trabajo técnico sobre equidad algorítmica, entre ellas la trampa del marco (evaluar el modelo aislado del sistema social que lo contiene) y la trampa del portabilidad (suponer que una solución justa en un contexto lo será en otro). Suresh y Guttag (2021) mostraron, de forma complementaria, que el daño no se origina en un punto único sino a lo largo de todo el ciclo de vida del aprendizaje automático: en el muestreo, en la agregación, en el aprendizaje, en la evaluación y en el despliegue.

Estas dos aportaciones justifican estructuralmente la matriz del ciclo de vida propuesta (Tabla 2) y la arquitectura en cuatro capas. Kellogg et al. (2020) documentaron que los algoritmos en el trabajo operan mediante seis mecanismos de control —dirección, evaluación y disciplina, en sus variantes restrictivas y habilitantes— que no son visibles en una auditoría técnica del modelo. De ahí que la capa relacional del modelo no sea un añadido humanista: es la única capa donde puede observarse si la calidad técnica se convierte en experiencia legítima o en una nueva forma de control opaco. La advertencia de Crawford (2021) y de Zuboff (2019) sobre la economía política de los sistemas de IA refuerza este punto: la unidad de análisis relevante incluye infraestructura, extracción de datos, relaciones laborales y asimetrías de poder que ningún indicador de precisión captura.

Cumplimiento regulatorio: suelo necesario, techo insuficiente

El Reglamento (UE) 2024/1689 (European Parliament and Council, 2024) representa un avance considerable al convertir la clasificación de usos en una obligación jurídica y al establecer prácticas prohibidas. El NIST AI Risk Management Framework (National Institute of Standards and Technology, 2023) aporta, por su parte, una estructura operativa —gobernar, mapear, medir y gestionar— que distribuye la gestión de riesgo a lo largo del ciclo de vida en lugar de concentrarla en una auditoría puntual. La norma ISO/IEC 42001 (ISO/IEC, 2023) añade la posibilidad de certificación de un sistema de gestión.

Sin embargo, el modelo IAcA sostiene que estos instrumentos, aun siendo necesarios, no agotan la responsabilidad, por tres razones que la literatura permite fundamentar. Primera: la clasificación regulatoria por nivel de riesgo deja fuera efectos relevantes sobre bienestar, cultura organizacional y calidad del trabajo que no se tipifican como daño jurídico. Segunda: existe un riesgo documentado de cumplimiento ritual, producir documentación sin modificar decisiones, que Raji et al. (2020) abordan al proponer un marco de auditoría interna capaz de generar consecuencias organizacionales reales, y que Kroll et al. (2017) anticipan al distinguir entre transparencia procedimental y responsabilidad efectiva. Tercera: el perfil de NIST para IA generativa (National Institute of Standards and Technology, 2024) evidencia que la velocidad de aparición de nuevas capacidades —confabulación, riesgos de integridad informativa, configuración humano-IA, coste ambiental— supera la cadencia de actualización normativa.

La implicación de diseño es explícita: la gobernanza sólo es efectiva si los hallazgos de evaluación pueden detener, rediseñar o retirar un sistema. Sin autoridad, presupuesto y consecuencias, la evaluación se convierte en burocracia que legitima sin corregir.

Justicia algorítmica: la imposibilidad técnica de eludir la decisión normativa

Una de las tensiones que el modelo asume de forma deliberada es la irreductibilidad de la justicia a una métrica. Barocas et al. (2023) y Binns (2018) han establecido que las distintas definiciones formales de equidad —paridad demográfica, igualdad de oportunidad, calibración— son matemáticamente incompatibles entre sí salvo en casos triviales, de modo que la elección de métrica es siempre una decisión normativa que no puede delegarse en el equipo técnico. Buolamwini y Gebu (2018) demostraron empíricamente que la evaluación agregada oculta disparidades severas que sólo emergen en el análisis desagregado por subgrupos interseccionales, y Noble (2018) mostró cómo los sistemas de recuperación de información reproducen y amplifican jerarquías sociales preexistentes.

Taylor (2017) desplaza el problema desde la equidad estadística hacia la justicia de datos, incorporando visibilidad, capacidad de impugnar la representación y no discriminación como dimensiones irreductibles a la métrica. Esta línea justifica dos exigencias del modelo que de otro modo parecerían excesivas: la evaluación desagregada obligatoria (P8) y el principio de no compensación, según el cual un beneficio agregado elevado no cancela una vulneración grave concentrada en un grupo específico.

A ello se suma un problema de precondition material. Sambasivan et al. (2021) documentaron la «cascada de datos»: los efectos compuestos y tardíos de una calidad de datos deficiente, sistemáticamente desvalorizada frente al trabajo de modelado. Gebru et al. (2021), con las hojas de datos, y Mitchell et al. (2019), con las tarjetas de modelo, ofrecen dispositivos concretos de documentación que el modelo incorpora como evidencia mínima en las fases de datos y validación. La justicia, en esta lectura, no se decide en el momento de la métrica sino mucho antes, en la construcción del conjunto de datos.

Explicabilidad, contestabilidad y distribución del poder

La explicabilidad ha sido tratada con frecuencia como problema técnico. Doshi-Velez y Kim (2017) advirtieron que la interpretabilidad carece de definición formal consensuada y que su evaluación depende del propósito y del destinatario. Ribeiro et al. (2016) aportaron métodos de explicación local, y Wachter et al. (2018) propusieron explicaciones contrafactuales que permiten a la persona afectada comprender qué habría tenido que ser distinto para obtener otro resultado, sin exigir la apertura del modelo.

El modelo IAcA adopta esta última orientación pero la subordina a una condición más exigente. Venkatasubramanian y Alfano (2020) fundamentan el derecho al *recourse* algorítmico en la autonomía y la capacidad de planificar: una explicación que no habilita acción no cumple función moral alguna. De ahí la distinción, central en la capa relacional, entre transparencia y contestabilidad. Pasquale (2015) y Citron y Pasquale (2014) mostraron que la opacidad de los sistemas de puntuación genera asimetrías de poder que la mera publicación de información no corrige: sin canales accesibles, plazos, revisión humana competente y capacidad real de rectificación, la transparencia se degrada en comunicación unilateral. Este es el fundamento de P4, que postula la contestabilidad como moderador del impacto de los errores algorítmicos sobre confianza y legitimidad.

Nissenbaum (2010) añade una precisión relevante para la capa de datos: la privacidad no se define por el secreto sino por la integridad contextual del flujo informativo. Un dato legítimamente recogido en un contexto puede vulnerar expectativas normativas al ser reutilizado en otro, lo que justifica evaluar la finalidad —y no sólo la base legal— en cada reutilización.

Agencia humana: contra la supervisión como ficción institucional

La crítica más incisiva que el modelo debe incorporar proviene de Green (2022), quien argumenta que las políticas que exigen supervisión humana de algoritmos gubernamentales suelen

fracasar porque asignan a las personas una tarea que la evidencia empírica muestra que no pueden desempeñar: identificar de forma fiable cuándo el sistema se equivoca. El resultado es doble y perverso: la supervisión no protege frente al error y, simultáneamente, legitima el despliegue y traslada la responsabilidad al operador individual.

Parasuraman y Riley (1997) habían establecido décadas antes la tríada de patologías de la automatización —uso, desuso y mal uso—, mostrando que tanto la sobreconfianza como la desconfianza sistemática degradan el desempeño conjunto. Amershi et al. (2019) y Shneiderman (2022) proponen, desde el diseño de interacción, que el objetivo no es maximizar la intervención humana sino calibrar la confianza mediante interfaces que comuniquen incertidumbre, límites y capacidad de corrección.

El modelo responde a esta crítica de tres maneras. Primera: define agencia no como presencia de un revisor, sino como conjunto de condiciones verificables —tiempo, información, formación, independencia y autoridad—, lo que la convierte en una variable medible y no en un requisito formal. Segunda: exige medir la eficacia de la supervisión, no su existencia, lo que convierte P3 en una proposición falsable. Tercera: establece que cuando la explicación resulta imposible y el impacto es alto, la respuesta legítima puede ser limitar o no desplegar el uso, en lugar de compensar la opacidad con una firma simbólica. Russell (2019) ofrece el fundamento de fondo: la delegación de objetivos a sistemas optimizadores sin mecanismos de deferencia y corrección es estructuralmente insegura, con independencia de la competencia del supervisor.

Bienestar sostenible como resultado, no como externalidad

La inclusión del bienestar entre los principios, y no entre los efectos deseables, constituye probablemente la decisión más distintiva del modelo frente a los marcos comparados en la Tabla 1. Su fundamento es triple. Deci y Ryan (2000) establecen que autonomía, competencia y vinculación son necesidades psicológicas cuya satisfacción determina la calidad de la motivación y del funcionamiento; Allan et al. (2019) documentan mediante metaanálisis la asociación del trabajo significativo con compromiso, satisfacción y desempeño; Edmondson (2019) muestra que la seguridad psicológica condiciona la disposición a reportar errores, precisamente la conducta de la que depende cualquier sistema de detección de incidentes algorítmicos.

De aquí se sigue una consecuencia poco advertida en la literatura de gobernanza: la capacidad de una organización para detectar daños de IA depende de su clima de bienestar. Si reportar un fallo es costoso, el sistema de monitorización queda ciego con independencia de su

sofisticación técnica. Esta relación fundamenta P6 —la intensificación del trabajo como mediador negativo entre automatización y bienestar— y P12.

El enfoque de capacidades de Sen (1999) ofrece el criterio evaluativo: la pregunta pertinente no es cuánta tarea se automatiza, sino qué puede realmente ser y hacer una persona después de la automatización. Si la eficiencia ganada se reinvierte íntegramente en volumen de producción, la IA habrá ampliado la capacidad de la organización y reducido la de las personas. De ahí P7 y el compromiso de reinversión parcial que aparece en el caso ilustrativo. La dimensión ambiental del bienestar sostenible, a su vez, se apoya en el análisis material de Crawford (2021) sobre el coste extractivo e infraestructural de los sistemas de IA, coste que rara vez aparece en la evaluación de un proyecto concreto.

La metáfora del alma: función, objeción y riesgo

La elección del término «alma» es el aspecto más expuesto a objeción del manuscrito y conviene defenderlo explícitamente. Su función es contrarrestar lo que van de Poel (2020) describe como reducción de valores a requisitos técnicos: cuando un valor se traduce en una especificación, tiende a perder el excedente normativo que lo hacía relevante. La metáfora preserva ese excedente al recordar que toda adopción tecnológica expresa una concepción de persona y de institución. En la tradición de Floridi (2013), se trata de reconocer que el diseño informacional es una forma de acción moral; en la de Pirson (2017), de subordinar la lógica económica a dignidad, bienestar y crecimiento humano; en la de Friedman et al. (2006), de incorporar valores de forma deliberada y trazable desde la fase de concepción.

El riesgo es simétrico y debe nombrarse: el lenguaje de la dignidad puede utilizarse para legitimar decisiones ya tomadas *-human-washing-*, del mismo modo que Hagendorff (2020) documenta el uso reputacional de las guías éticas. Las defensas que el modelo opone son estructurales y no retóricas: trazabilidad de las decisiones éticas (P11), participación de personas afectadas con capacidad de modificar el resultado y no sólo de ser consultadas (P2), auditoría con consecuencias siguiendo a Raji y Buolamwini (2019) y Raji et al. (2020), y existencia efectiva de decisiones de pausa y retirada. Conviene subrayar que el modelo no atribuye interioridad a la máquina. Sostiene lo contrario: la tecnología carece de finalidad moral propia, y por ello la responsabilidad permanece íntegramente en las personas e instituciones que la desarrollan, adquieren y utilizan.

Implicaciones para la dirección y el gobierno

Las implicaciones prácticas del análisis anterior pueden formularse por nivel de responsabilidad.

Consejos de administración y órganos de gobierno universitario. La IA debe tratarse como asunto estratégico y de riesgo, no exclusivamente tecnológico. Ello exige un inventario de sistemas en uso, identificación de las decisiones sobre las que influyen, asignación nominal de responsabilidad y reporte periódico de incidentes, siguiendo la lógica de la función *govern* del marco de NIST (National Institute of Standards and Technology, 2023).

Alta dirección. La coherencia entre capas (P9) depende de los incentivos. Si se premia únicamente la velocidad de implantación, la evaluación ética quedará subordinada por diseño. Los responsables de gobernanza necesitan autoridad formal para pausar proyectos, no sólo para emitir recomendaciones.

Equipos técnicos. La documentación de supuestos, límites y cambios —mediante hojas de datos y tarjetas de modelo (Gebru et al., 2021; Mitchell et al., 2019)— debe entenderse como parte de la excelencia técnica y no como carga administrativa. La capacidad de explicar y mantener un sistema es tan relevante como la de construirlo.

Funciones transversales. La participación de recursos humanos, asesoría jurídica, cumplimiento, seguridad, sostenibilidad y personas usuarias mejora la información disponible sin diluir la responsabilidad, siempre que los roles estén formalmente asignados (ISO/IEC, 2023).

Gestión de proveedores. La dependencia tecnológica constituye un riesgo de gobierno. La evaluación debe abarcar documentación, derecho de auditoría, seguridad, condiciones de tratamiento de datos, accesibilidad, notificación de cambios de modelo y plan de salida.

Formación. La alfabetización crítica debe alcanzar a toda la plantilla con competencias diferenciadas por rol (UNESCO, 2023). Saber utilizar un sistema no equivale a saber gobernarlo, y la sobreconfianza y la infrautilización son fallos simétricos (P10).

Limitaciones

La primera limitación es de naturaleza epistémica. El modelo es una construcción teórico-propositiva: su coherencia interna y su fundamentación en la literatura no equivalen a evidencia de eficacia. Ninguna de las doce proposiciones ha sido contrastada empíricamente, y la escala IAca-36 carece por completo de validación psicométrica, por lo que no debe emplearse como instrumento de medida hasta completar el protocolo descrito.

La segunda es de parsimonia. La amplitud del modelo —seis principios, cuatro capas, siete fases, seis dimensiones de auditoría— facilita la cobertura conceptual pero dificulta la aplicación en organizaciones con recursos limitados y plantea un riesgo real de solapamiento entre dimensiones. La investigación futura deberá determinar si las seis dimensiones constituyen factores diferenciados o una capacidad organizacional de segundo orden.

La tercera es cultural. Dignidad, autonomía, privacidad y bienestar admiten interpretaciones divergentes entre contextos. La adaptación cultural es legítima, pero no puede convertirse en argumento para aceptar discriminación o ausencia de vías de recurso; el equilibrio entre universalidad de derechos y legitimidad local permanece sin resolver en el modelo.

La cuarta es la velocidad del cambio tecnológico. Nuevas arquitecturas y modalidades de uso modifican el perfil de riesgo con mayor rapidez que cualquier marco de gobierno, como ilustra la necesidad de un perfil específico para IA generativa (National Institute of Standards and Technology, 2024).

La quinta es la disponibilidad de datos. Numerosos impactos sociales y laborales no se registran de forma sistemática, por lo que la evaluación deberá combinar métricas cuantitativas con entrevistas, observación, análisis de reclamaciones y registros de incidentes.

Finalmente, debe explicitarse que el caso presentado es hipotético y cumple una función ilustrativa, no probatoria. La trayectoria profesional del autor se ha empleado como fuente de problematización y no como evidencia.

Conclusiones

La inteligencia artificial no llega a una organización como una herramienta neutral. Reconfigura decisiones, conocimiento, poder y experiencia. Por ello, su adopción no puede delegarse exclusivamente en equipos técnicos ni evaluarse solo por precisión.

El modelo Inteligencia Artificial con Alma propone seis principios, cuatro capas y un ciclo 4A para convertir valores en práctica. Su idea central es sencilla: la tecnología debe ampliar capacidades humanas sin convertir la dignidad, la justicia y el bienestar en variables residuales.

Una IA con alma no es una máquina que siente. Es una institución que asume responsabilidad. Es la capacidad humana de decidir para qué se utiliza la tecnología, qué límites se establecen, quién participa, cómo se corrigen errores y cuándo se renuncia a un uso.

El futuro digital no será más humano por defecto. Será humano si se diseña, gobierna y evalúa con conciencia. Transformar con alma significa que cada avance técnico debe responder a una pregunta moral y organizacional: ¿esta tecnología nos ayuda a ser más capaces, más libres, más justos y más responsables?

Agenda de investigación 2026-2035

La consolidación de una inteligencia artificial humanista requiere una agenda acumulativa y abierta. El primer eje consiste en validar la escala IAca-36 en organizaciones, universidades, servicios públicos y sistemas de salud. La investigación deberá comprobar su estructura, estabilidad, sensibilidad al cambio y relación con resultados observables. También será necesario determinar si las seis dimensiones funcionan como factores diferenciados o como una capacidad organizacional de segundo orden.

El segundo eje se centra en la supervisión humana. Es preciso estudiar qué combinaciones de explicación, experiencia, tiempo, autoridad y responsabilidad permiten una intervención efectiva. Experimentos controlados y estudios de campo pueden comparar supervisión nominal, revisión activa y decisión compartida. El objetivo no es maximizar la intervención humana en todos los casos, sino identificar cuándo añade valor y evita dependencia.

El tercer eje aborda bienestar y calidad del trabajo. Los estudios longitudinales deberían analizar si los asistentes de IA reducen carga o simplemente aumentan expectativas de producción; si amplían creatividad o estandarizan; y si mejoran aprendizaje o provocan pérdida de habilidades. Estas preguntas requieren indicadores subjetivos, conductuales y organizacionales.

El cuarto eje estudia justicia y participación. Se necesitan diseños que comparen proyectos con y sin co-diseño, y que midan si la participación modifica decisiones, no solo percepciones. También debe investigarse cómo incluir a grupos con menor poder, cómo compensar el tiempo de participación y cómo resolver conflictos entre intereses legítimos.

El quinto eje examina proveedores y cadenas de valor. La evidencia sobre transparencia contractual, actualizaciones, lock-in y reparto de responsabilidades sigue siendo limitada. Estudios comparativos pueden identificar cláusulas y prácticas que mejoran trazabilidad y capacidad de salida.

El sexto eje se refiere a IA generativa y conocimiento. Será necesario estudiar cómo cambian la autoría, la memoria institucional, el razonamiento y la confianza cuando los sistemas participan en producción de textos, análisis y decisiones. La calidad no debe medirse solo por

similitud con respuestas humanas, sino por su contribución a comprensión, deliberación y aprendizaje.

El séptimo eje es intercultural. El modelo debe contrastarse en contextos con distintas concepciones de autonomía, comunidad, privacidad y autoridad. La adaptación cultural no puede convertirse en excusa para aceptar discriminación o ausencia de recurso; debe ayudar a traducir principios universales en prácticas legítimas.

El octavo eje es ambiental. Se requiere vincular beneficios de uso con consumo, infraestructura y vida útil de dispositivos. La evaluación debe evitar tanto ignorar el coste material como atribuir estimaciones imprecisas. La proporcionalidad computacional puede convertirse en una línea específica de diseño responsable.

Finalmente, la investigación debería explorar si la madurez IAcA predice innovación sostenible, confianza, reducción de incidentes y legitimidad. Una hipótesis relevante es que la ética aplicada no ralentiza necesariamente la transformación digital: puede evitar proyectos fallidos, mejorar adopción y orientar recursos hacia usos socialmente valiosos.

Declaraciones

- Conflicto de intereses: el autor declara que no existe conflicto de intereses.
- Financiación: el estudio no recibió financiación externa específica.
- Disponibilidad de datos: no aplica, por tratarse de un estudio teórico-propositivo.
- Uso de inteligencia artificial: las herramientas digitales, en su caso, se utilizaron como apoyo editorial. La conceptualización, interpretación y responsabilidad final corresponden al autor.
- Originalidad: el manuscrito es original y no se encuentra sometido simultáneamente a otra revista. Las referencias incluidas en la lista final se encuentran citadas expresamente en el cuerpo del manuscrito.

Referencias

- Allan, B. A., Batz-Barbarich, C., Sterling, H. M., & Tay, L. (2019). Outcomes of meaningful work: A meta-analysis. *Journal of Management Studies*, 56(3), 500-528.
- Amershi, S., Weld, D., Vorvoreanu, M., et al. (2019). Guidelines for human-AI interaction. *Proceedings of CHI 2019*.
- Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning. fairmlbook.org.

- Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of FAT*.
- Buolamwini, J., & Gebru, T. (2018). Gender shades. *Proceedings of FAT*.
- Citron, D. K., & Pasquale, F. (2014). The scored society. *Washington Law Review*, 89, 1-33.
- Crawford, K. (2021). *Atlas of AI*. Yale University Press.
- Deci, E. L., & Ryan, R. M. (2000). The what and why of goal pursuits. *Psychological Inquiry*, 11(4), 227-268.
- Dignum, V. (2019). *Responsible artificial intelligence*. Springer.
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*.
- Edmondson, A. C. (2019). *The fearless organization*. Wiley.
- European Commission High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI*.
- European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence.
- Floridi, L. (2013). *The ethics of information*. Oxford University Press.
- Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1).
- Freeman, R. E. (1984). *Strategic management: A stakeholder approach*. Pitman.
- Friedman, B., Kahn, P. H., & Borning, A. (2006). Value sensitive design and information systems. In *Human-computer interaction and management information systems*.
- Gebru, T., Morgenstern, J., Vecchione, B., et al. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92.
- Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45.
- Hagendorff, T. (2020). The ethics of AI ethics. *Minds and Machines*, 30, 99-120.
- IEEE. (2019). *Ethically aligned design*. IEEE.
- ISO/IEC. (2023). *ISO/IEC 42001:2023 Information technology - Artificial intelligence - Management system*.
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, 389-399.

- Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work. *Academy of Management Annals*, 14(1), 366-410.
- Kroll, J. A., Huey, J., Barocas, S., et al. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165, 633-705.
- Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making. *Philosophy & Technology*, 31, 611-627.
- Mitchell, M., Wu, S., Zaldivar, A., et al. (2019). Model cards for model reporting. *Proceedings of FAT*.
- Mittelstadt, B. D. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1, 501-507.
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms. *Big Data & Society*, 3(2).
- National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1.
- National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1.
- Nissenbaum, H. (2010). *Privacy in context*. Stanford University Press.
- Noble, S. U. (2018). *Algorithms of oppression*. NYU Press.
- OECD. (2019). *Recommendation of the Council on Artificial Intelligence*.
- O'Neil, C. (2016). *Weapons of math destruction*. Crown.
- Parasuraman, R., & Riley, V. (1997). Humans and automation. *Human Factors*, 39(2), 230-253.
- Pasquale, F. (2015). *The black box society*. Harvard University Press.
- Pirson, M. (2017). *Humanistic management*. Cambridge University Press.
- Raji, I. D., Smart, A., White, R. N., et al. (2020). Closing the AI accountability gap. *Proceedings of FAT*.
- Raji, I. D., & Buolamwini, J. (2019). Actionable auditing. *Proceedings of AIES*.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? *Proceedings of KDD*.
- Russell, S. (2019). *Human compatible*. Viking.
- Sambasivan, N., Kapania, S., Highfill, H., et al. (2021). Everyone wants to do the model work, not the data work. *Proceedings of CHI*.

- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of FAT.*
- Sen, A. (1999). *Development as freedom*. Oxford University Press.
- Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.
- Suresh, H., & Gutttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Proceedings of EAAMO*.
- Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, 361(6404), 751-752.
- Taylor, L. (2017). What is data justice? *Big Data & Society*, 4(2).
- UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*.
- UNESCO. (2023). *Guidance for generative AI in education and research*.
- van de Poel, I. (2020). Embedding values in artificial intelligence systems. *Minds and Machines*, 30, 385-409.
- Venkatasubramanian, S., & Alfano, M. (2020). The philosophical basis of algorithmic recourse. *Proceedings of FAT*.
- Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box. *Harvard Journal of Law & Technology*, 31(2), 841-887.
- World Health Organization. (2021). *Ethics and governance of artificial intelligence for health*.
- Zuboff, S. (2019). *The age of surveillance capitalism*. PublicAffairs.