



ARTÍCULO

Gobernanza ético-tecnológica de la inteligencia artificial generativa: hacia un marco de integridad centrada en el humano en contextos de innovación digital

Ethical-technological governance of generative artificial intelligence: toward a human-centered integrity framework in contexts of digital innovation

Wacner Galindez 

Universidad Nacional Experimental de los Llanos Occidentales Ezequiel Zamora, Barinas, Venezuela

Autor de Correspondencia: wacner@unellez.edu.ve

Enviado: 25 Abril 2026 Aceptado: 12 Julio 2026 Publicado en línea: 15 Julio 2026

Resumen

La expansión de la inteligencia artificial generativa ha puesto en evidencia que la abundancia de principios éticos contrasta con la escasez de mecanismos verificables para aplicarlos. Ante esta brecha, el estudio diseña y evalúa un marco de integridad ético-tecnológica que orienta la innovación digital hacia la transparencia algorítmica y la salvaguarda de derechos fundamentales sin inhibir el desarrollo técnico. Siguiendo el paradigma de la ciencia del diseño, el marco se construyó a partir de cinco requisitos derivados de la teoría y de un análisis documental del Reglamento (UE) 2024/1689, del marco del NIST y de las disposiciones chinas sobre IA generativa, codificados mediante una matriz de ocho indicadores. Los resultados muestran que ninguno de los tres regímenes garantiza plenamente el veto humano vinculante ni la participación de los afectados. En respuesta, el marco propone cuatro ejes, que comprenden la auditabilidad, la responsabilidad proactiva, la mitigación participativa de sesgos y un protocolo de escalamiento segmentado por riesgo. Su demostración en un servicio público de protección social indica que el marco es operable y convierte la ética en un activo estratégico al generar evidencias auditables.

Palabras clave: Inteligencia artificial generativa, gobernanza algorítmica, jurisprudencia tecnológica, integridad digital, ética centrada en el humano, transparencia algorítmica

Abstract

The expansion of generative artificial intelligence has revealed that the abundance of ethical principles contrasts with the scarcity of verifiable mechanisms to apply them. In response to this gap, the study designs and evaluates an ethical-technological integrity framework that steers digital innovation toward algorithmic transparency and the safeguarding of fundamental rights without inhibiting technical development. Following the design science paradigm, the framework was built from five theory-derived requirements and a document analysis of Regulation (EU) 2024/1689, the NIST framework, and China's provisions on generative AI, coded through a matrix of eight indicators. The results show that none of the three regimes fully guarantees binding human veto or the participation of affected groups. In response, the framework proposes four axes, namely auditability, proactive accountability, participatory bias mitigation, and a risk-segmented escalation protocol. Its demonstration in a social protection public service indicates that the framework is operable and turns ethics into a strategic asset by generating auditable evidence.

Keywords: Generative artificial intelligence, algorithmic governance, technological jurisprudence, digital integrity, human-centered ethics, algorithmic transparency

Este es un artículo de acceso abierto y se distribuye bajo los términos de la Licencia Creative Commons Atribución-NoComercial-SinDerivadas 4.0 Internacional (CC BY-NC-ND 4.0).

Citar como: W. Galindez, "Gobernanza ético-tecnológica de la inteligencia artificial generativa: hacia un marco de integridad centrada en el humano en contextos de innovación digital," *CISAI*, vol. 2, no. 2, pp. 63–81, Julio 2026, doi: 10.64439/cisai.v2i2.46.

1. Introducción

La proliferación de los modelos de lenguaje extensos ha reconfigurado la arquitectura de la innovación digital durante el bienio 2022-2024, en un proceso impulsado por el crecimiento sostenido de los recursos computacionales destinados al entrenamiento, cuya tasa de duplicación se ha acortado de manera notable en la última década [1]. Este escalamiento ha dado lugar a capacidades que emergen de forma abrupta y difícil de anticipar, de modo que los propios desarrolladores reconocen que el comportamiento de estos sistemas combina una predictibilidad agregada con sorpresas específicas en tareas concretas [2]. Como consecuencia, su adopción en sectores estratégicos, desde los servicios jurídicos hasta la educación y la salud, ha avanzado con mayor rapidez que la comprensión de sus implicaciones organizacionales, éticas y regulatorias [3]. Dicha expansión, sin embargo, no ha estado exenta de externalidades negativas, entre las que destacan la generación de contenido falso con apariencia verosímil [4], la reproducción de estereotipos presentes en los corpus de entrenamiento [5] y un conjunto más amplio de riesgos que abarca la desinformación, la exposición de datos privados y los daños derivados de la interacción entre humanos y sistemas conversacionales [6].

En este escenario se observa una asimetría crítica entre la velocidad de iteración técnica de la inteligencia artificial generativa (IAG) y la capacidad de respuesta de los marcos regulatorios. Smuha [7] documentó cómo la competencia geopolítica transitó de una carrera por la IA hacia una carrera por su regulación, en la que la Unión Europea, Estados Unidos y China despliegan arquitecturas normativas divergentes que difícilmente logran sincronizarse con los ciclos de desarrollo del software. Esta divergencia se aprecia también en la fragmentación de los modos de gobernanza, que combinan comités éticos, autorregulación industrial, estándares técnicos y legislación con efectos extraterritoriales sin que exista una lógica común de coordinación [8]. En el caso europeo, Veale y Zuiderveen [9] advirtieron, al analizar el borrador del Reglamento de IA, que la categorización por niveles de riesgo presentaba lagunas respecto a los modelos de propósito general, cuya versatilidad dificulta una clasificación estática. Con la adopción del Reglamento (UE) 2024/1689, esta tensión no desapareció, puesto que la noción de confiabilidad que inspira la norma tiende a confundirse con la aceptabilidad del riesgo, lo que desplaza hacia los proveedores decisiones que tienen un marcado carácter político [10].

Frente a esta problemática, la literatura especializada ha identificado la ausencia de marcos integrales que articulen jurisprudencia tecnológica, ética por diseño y gobernanza algorítmica en un sistema coherente aplicable a la IAG. Jobin et al. [11] mapearon ochenta y cuatro documentos de ética de la IA y constataron que la mayoría converge en principios como la transparencia, la justicia y la no maleficencia, aunque difiere en su interpretación y rara vez especifica cómo aplicarlos. En la misma dirección, Mittelstadt [12] sostuvo que los principios por sí solos no garantizan una IA ética mientras no se vinculen con mecanismos de rendición de cuentas y procesos de verificación, y Hagendorff [13] mostró que las directrices vigentes carecen, en su mayoría, de mecanismos de refuerzo capaces de modificar las prácticas de desarrollo. Esta distancia entre la abstracción normativa y la concreción técnica, que Morley et al. [14] caracterizaron como el tránsito pendiente del qué al cómo, constituye el vacío específico que aborda el presente estudio.

En consecuencia, el objetivo general de la investigación consiste en diseñar y evaluar un marco de integridad ético-tecnológica para la IAG que canalice la innovación digital bajo principios de jurisprudencia tecnológica, transparencia algorítmica y salvaguarda de derechos fundamentales, sin generar efectos inhibitorios sobre el desarrollo técnico. Este objetivo se desagrega en cuatro propósitos específicos, que consisten en sistematizar la evolución de la responsabilidad algorítmica frente a modelos con autonomía operativa, identificar los mecanismos de ética por diseño aplicables al ciclo de vida de la IAG, construir una matriz analítica que integre transparencia, responsabilidad, mitigación de sesgos e interacción humano-IA, y discutir, con carácter exploratorio, las implicaciones del marco sobre la competitividad y la confianza en el mercado digital. Los tres primeros propósitos se traducen en requisitos de diseño evaluables, mientras que el cuarto se aborda de forma argumentativa en la discusión y queda abierto a una validación empírica posterior. La distinción entre principios orientadores y prácticas implementables, propuesta por Floridi y Cowls [15], proporciona el fundamento conceptual de esta operacionalización, que se aborda como una investigación original en ciencia del diseño [16], puesto que el marco se concibe como un artefacto que se construye, se demuestra y se evalúa frente a requisitos explícitos, y no como una síntesis de la literatura

existente.

A partir de estos propósitos, el trabajo realiza tres contribuciones. La primera es una matriz de evaluación de ocho indicadores que permite comparar regímenes regulatorios heterogéneos con un lenguaje común y que, aplicada a los tres marcos de referencia, revela vacíos compartidos en el veto humano vinculante y en la participación de los grupos afectados. La segunda es un marco de integridad de cuatro ejes que traduce cada vacío identificado en mecanismos operativos y en evidencias verificables por terceros, y la tercera es un protocolo de escalamiento humano segmentado por riesgo, formalizado como algoritmo y como diagrama de secuencia, que busca conciliar la supervisión efectiva con la escala de operación propia de estos sistemas, cuya aplicabilidad se demuestra en un escenario de servicios públicos y se evalúa frente a cinco requisitos de diseño. El artículo se organiza de modo que la Sección 2 examina los trabajos relacionados, la Sección 3 expone los fundamentos teóricos, la Sección 4 describe el diseño metodológico, la Sección 5 presenta los resultados y el marco propuesto, y las Secciones 6 y 7 discuten sus implicaciones y conclusiones.

2. Trabajos relacionados

La literatura sobre gobernanza de la IA ha avanzado por varias vías que, aunque complementarias, rara vez convergen en una misma propuesta. Una primera vía se ha concentrado en traducir principios éticos en herramientas prácticas, y en ella Morley et al. [14] revisaron los instrumentos disponibles para cada fase del ciclo de vida del aprendizaje automático, concluyendo que la mayoría se encuentra en estadios incipientes de validación y que su distribución entre fases es desigual. Este diagnóstico, si bien valioso para orientar a los desarrolladores, precede a la irrupción de los modelos generativos y no contempla los desafíos asociados a sistemas de propósito general cuyo uso final escapa al control de quien los entrena.

Una segunda vía, de orientación jurídica, ha examinado cómo los ordenamientos existentes responden a la IAG. Hacker et al. [17] analizaron la regulación de los grandes modelos generativos en el contexto europeo y propusieron distinguir entre desarrolladores, desplegados y usuarios para distribuir obligaciones de forma proporcionada, mientras que Laux et al. [10] cuestionaron que el Reglamento europeo equipare la confiabilidad con la aceptabilidad del riesgo. Por su parte, Veale et al. [8] ampliaron el foco hacia la gobernanza global, identificando la coexistencia de modalidades regulatorias que responden a racionalidades distintas. Estos trabajos aportan un análisis jurídico riguroso, aunque su alcance se detiene en la crítica normativa y no desciende hasta los mecanismos técnicos que harían exigibles las obligaciones propuestas.

Desde una tercera vía, centrada en la auditoría, Mökander et al. [18] formularon un enfoque de tres capas, que distingue auditorías de gobernanza, de modelo y de aplicación, como mecanismo para identificar y gestionar los riesgos de los modelos de lenguaje. Esta propuesta representa un avance operativo significativo, pero su énfasis recae en la verificación y no en la asignación de responsabilidades ni en el diseño de la supervisión humana durante la operación. De forma complementaria, Novelli et al. [19] precisaron el concepto de rendición de cuentas en la IA y sus componentes relacionales, en tanto que Papagiannidis et al. [20] organizaron la gobernanza responsable en dimensiones estructurales, relacionales y procedimentales desde la perspectiva de los sistemas de información. Ambos aportes enriquecen el vocabulario conceptual del campo, aunque no se orientan específicamente a la IAG ni ofrecen una comparación entre regímenes regulatorios. La Tabla 1 sintetiza este análisis a partir de cinco criterios que, en conjunto, definen el espacio que pretende ocupar el presente estudio. Como puede apreciarse, ninguno de los trabajos examinados combina la especificidad sobre la IAG, el anclaje jurídico, la operacionalización técnica, la comparación entre regímenes y un mecanismo de supervisión humana vinculante, lo que justifica la necesidad de un marco que articule estas dimensiones en una misma estructura.

3. Marco teórico: la tríada de la gobernanza

Los antecedentes revisados exponen que la gobernanza de la IAG descansa sobre tres pilares cuya interacción, más que su consideración aislada, determina la eficacia de cualquier marco. Por esta razón, la presente sección examina sucesivamente la jurisprudencia tecnológica, la ética por diseño y la relación entre gobernanza e innovación digital, para después contrastar las arquitecturas regulatorias que concretan estos pilares en tres jurisdicciones.

Tabla 1. Comparación de trabajos relacionados según los criterios que definen el aporte del estudio

Trabajo	Específico IAG	Anclaje jurídico	Operacionalización	Comparación regulatoria	Veto humano
Morley et al. [14]	–	–	✓	–	–
Hacker et al. [17]	✓	✓	◦	–	–
Veale et al. [8]	–	✓	–	✓	–
Laux et al. [10]	–	✓	–	–	–
Novelli et al. [19]	–	◦	◦	–	–
Mökander et al. [18]	✓	◦	✓	–	◦
Papagiannidis et al. [20]	–	–	◦	–	–
Presente estudio	✓	✓	✓	✓	✓

✓ abordado de forma central ◦ abordado parcialmente – no abordado

3.1 Jurisprudencia tecnológica e IAG

Conviene precisar, en primer lugar, el alcance con el que se emplea esta expresión. En la tradición jurídica hispanohablante, el término jurisprudencia suele designar el conjunto de resoluciones judiciales que fijan criterios interpretativos. En este estudio, en cambio, la expresión jurisprudencia tecnológica se utiliza en un sentido más amplio, próximo al de *technology law*, para referirse al conjunto de normas, doctrina y criterios interpretativos que regulan la atribución de responsabilidad y la protección de derechos frente a las tecnologías digitales. Por consiguiente, el análisis no se basa en sentencias, sino en textos normativos y en la doctrina que los interpreta, y el pilar se entiende como la dimensión jurídica de la gobernanza de la IAG.

La evolución del derecho frente a lo intangible ha transitado desde una concepción de la responsabilidad centrada en el programador hacia un paradigma que debe atender la autonomía operativa de los modelos de aprendizaje profundo. Matthias [21] anticipó que los sistemas capaces de aprender generarían una brecha de responsabilidad, puesto que ni el fabricante ni el operador controlan plenamente su comportamiento, y Santoni de Sio y Mecacci [22] precisaron posteriormente que dicha brecha se manifiesta en cuatro formas distintas, relativas a la culpabilidad, la rendición de cuentas moral, la rendición de cuentas pública y la responsabilidad activa. En el ámbito de la protección de datos, Wachter et al. [23] argumentaron que el Reglamento General de Protección de Datos no garantiza un derecho general a la explicación de la lógica interna de los algoritmos, sino únicamente a información sobre la existencia del procesamiento automatizado y sus consecuencias, mientras que Kaminski [24] ofreció una lectura más amplia, según la cual el reglamento sí impone una obligación de explicación significativa que opera en combinación con otras salvaguardas.

Esta insuficiencia se magnifica en la IAG, donde la distancia entre los datos de entrenamiento, la arquitectura del modelo y el contenido generado dificulta la atribución causal de los daños. Bryson [25] propuso que los sistemas inteligentes pueden ser objeto de consideración sin que ello implique concederles agencia moral, lo que exige mantener la cadena de responsabilidad en los actores humanos del ecosistema, es decir, en desarrolladores, desplegados y usuarios finales. A esta perspectiva se suma la constatación de Cath et al. [26], quienes compararon los enfoques de Estados Unidos, la Unión Europea y el Reino Unido y encontraron que la fragmentación jurisdiccional genera zonas grises cuando los modelos operan de manera transnacional. De ahí que la cuestión jurídica central ya no sea únicamente quién responde, sino cómo distribuir la carga de la prueba cuando un modelo fundacional produce resultados difamatorios o discriminatorios en contextos de uso no previstos por sus desarrolladores originales, una incertidumbre que Hacker et al. [17] identifican como un factor de riesgo sistémico para la inversión en innovación.

3.2 Ética por diseño

El segundo pilar se refiere a la integración de principios éticos en el ciclo de vida del desarrollo. El marco AI4People [27] sintetizó las propuestas previas en cinco principios, beneficencia, no maleficencia, autonomía, justicia y explicabilidad, que luego inspiraron los requisitos de una IA confiable adoptados en el ámbito europeo. No obstante, Schiff et al. [28] observaron, al revisar documentos de los sectores público, privado y no gubernamental, que existe una convergencia retórica acompañada de una divergencia sustancial en la implementación, y Floridi [29] advirtió sobre riesgos como el lavado ético, que consiste en proyectar una imagen de responsabilidad sin cambios efectivos en las prácticas. Cuando estas consideraciones se trasladan a la IAG, surgen limitaciones estructurales adicionales. Los mecanismos de alineamiento basados en retroalimentación humana constituyen, en sentido estricto, una forma de ética por diseño, aunque su eficacia depende de la representatividad de los juicios incorporados y de la concepción de valor que se adopte en cada contexto comunicativo [30]. Además, la aparición de capacidades no anticipadas durante el entrenamiento [2] implica que ningún conjunto finito de reglas codificadas puede prever todos los comportamientos del sistema. Por consiguiente, la ética por diseño en la IAG exige una reconceptualización que trascienda la codificación de reglas y aborde la dinámica de aprendizaje y despliegue continuo del modelo, lo que conduce de manera natural a la relación entre gobernanza e innovación.

3.3 Innovación digital y gobernanza

Continuando con lo anterior, la tercera dimensión examina si la gobernanza actúa como freno o como condición habilitante de la innovación. Zuboff [31] caracterizó la lógica del capitalismo de la vigilancia como un modelo en el que la acumulación de datos se antepone a los derechos de las personas afectadas, lo que ilustra cómo la ausencia de reglas puede orientar la innovación hacia la extracción antes que hacia la creación de valor compartido. Birhane [32] complementó esta crítica al mostrar que los sistemas desplegados sin atención a su contexto relacional perpetúan injusticias que erosionan la legitimidad social de la tecnología. En sentido inverso, la literatura sobre gobernanza organizacional sostiene que las estructuras y procesos de gobernanza responsable permiten a las organizaciones capturar el valor de la IA mientras gestionan sus riesgos [20], y que dicha gobernanza requiere todavía una agenda de investigación que conecte los principios con las prácticas empresariales [33]. Esta visión debe matizarse, sin embargo, con el reconocimiento de tensiones inherentes. Whittaker [34] documentó cómo la industria tecnológica ha capturado parcialmente la investigación en ética de la IA, con el efecto de neutralizar la presión regulatoria, y Taddeo et al. [35] advirtieron que confiar en la IA para fines de seguridad constituye un arma de doble filo, dado que los mismos sistemas pueden ser atacados o manipulados. Por tanto, cualquier marco de gobernanza que aspire a funcionar como infraestructura de confianza debe incorporar mecanismos de verificación independiente, pues solo así se evita que la ética se reduzca a una estrategia de relaciones públicas.

4. Metodología

El estudio se enmarca en la investigación en ciencia del diseño, un paradigma orientado a crear y evaluar artefactos que resuelven problemas relevantes y que, al mismo tiempo, aportan conocimiento prescriptivo sobre cómo resolverlos [16, 36]. Bajo este enfoque, el marco de integridad constituye el artefacto de la investigación, de modo que su contribución no reside en sintetizar lo publicado, sino en proponer una solución nueva, fundamentada en teorías de base y en evidencia documental primaria, cuya utilidad se demuestra y evalúa de manera explícita. Siguiendo la metodología de Peffers et al. [37], el proceso se organizó en seis actividades que la Figura 1 vincula con las secciones del artículo, y en las que la literatura cumple la función de teoría de base para derivar los requisitos de diseño, mientras que los textos normativos constituyen los datos primarios del análisis.

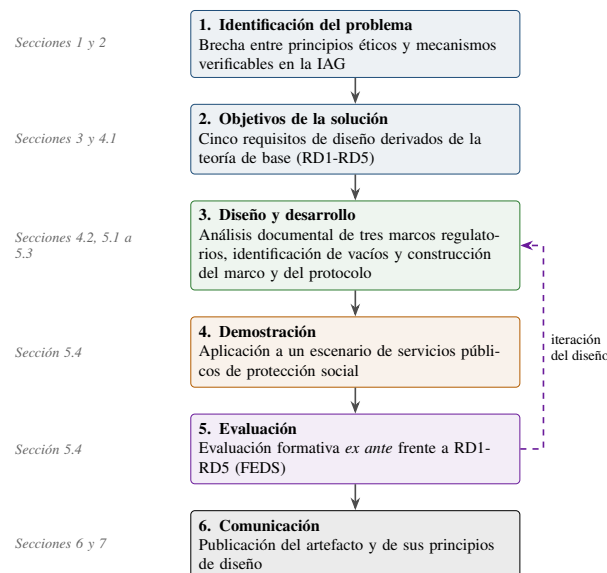


Figura 1. Proceso de investigación en ciencia del diseño seguido en el estudio, adaptado de Peffers et al. [37].

4.1 Identificación del problema y requisitos de diseño

El problema que motiva el artefacto, descrito en las Secciones 1 y 2, consiste en que la gobernanza de la IAG dispone de principios ampliamente compartidos, pero carece de mecanismos que permitan verificar su cumplimiento y asignar responsabilidades cuando fallan [12, 14]. Con el fin de transformar este problema en objetivos de solución, se derivaron de la teoría de base cinco requisitos de diseño que cualquier respuesta satisfactoria debería cumplir y que, a su vez, sirven como criterios de evaluación del artefacto. La Tabla 2 presenta estos requisitos junto con su fundamento teórico.

Tabla 2. Requisitos de diseño del artefacto derivados de la teoría de base

Cód.	Requisito	Enunciado	Fundamento
RD1	Auditabilidad	Los mecanismos deben producir evidencias verificables por terceros, más allá de la divulgación de información	[38, 12]
RD2	Asignación <i>ex ante</i> de responsabilidad	Los roles y obligaciones deben definirse antes del despliegue, incluso frente a capacidades no anticipadas	[21, 22, 2]
RD3	Participación de afectados	Los grupos afectados deben intervenir en la definición del daño y en la validación de las métricas de equidad	[39, 40]
RD4	Supervisión humana efectiva	El operador debe contar con autoridad, competencia y una carga de trabajo compatibles con el control real	[41, 42]
RD5	Proporcionalidad	La intensidad de los controles debe ajustarse al riesgo para no inhibir la innovación	[43, 20]

4.2 Análisis documental de los marcos regulatorios

La actividad de diseño se apoyó en un análisis documental cualitativo [44] de tres marcos regulatorios, que constituyen los datos primarios del estudio. Se analizaron el Reglamento (UE) 2024/1689 [45], el marco de gestión de riesgos de IA del NIST en su versión 1.0 de 2023 [46] y el conjunto de disposiciones chinas sobre algoritmos de recomendación, síntesis profunda y servicios de IA generativa adoptadas entre 2021 y 2023. A diferencia de los otros dos instrumentos, el marco del NIST es de adopción voluntaria, y se seleccionó porque constituye la referencia técnica federal más difundida en Estados Unidos ante la ausencia de una ley federal general sobre IA. Esta asimetría se asume de manera deliberada, puesto que la escala de valoración distingue entre previsiones recomendadas y vinculantes, de modo que la menor puntuación estadounidense refleja en parte el carácter voluntario del instrumento y no solo su contenido. Por la misma razón, el perfil de IA generativa publicado por el NIST en 2024 como complemento del marco [47], que también es voluntario, no alteraría el nivel máximo alcanzable por Estados Unidos en ningún indicador, aunque sí podría enriquecer el contenido de las previsiones recomendadas, aspecto que se recoge entre las

limitaciones del estudio. Las disposiciones chinas se analizaron a partir de las traducciones no oficiales al inglés publicadas por China Law Translate, sin cotejo con el texto original en chino, circunstancia que se recoge entre las limitaciones del estudio.

El instrumento de análisis fue una matriz construida *ad hoc* a partir de los cuatro ejes del marco, cada uno desagregado en dos indicadores alineados con los requisitos de diseño, como se detalla en la Tabla 3. Cada indicador se valoró con una escala ordinal de tres niveles, en la que el valor 0 indica ausencia de previsión normativa, el valor 1 indica una previsión parcial, recomendada o limitada a ciertos actores, y el valor 2 indica una previsión plena, vinculante y verificable. De este modo, cada eje admite una puntuación entre 0 y 4 y cada marco una puntuación global entre 0 y 16.

Tabla 3. Operacionalización de la matriz de análisis documental

Eje	Cód.	Indicador	Criterio para el nivel pleno (2)
I. Transparencia y explicabilidad	T1	Documentación técnica	Documentación obligatoria del modelo, de los datos y de sus versiones, accesible a la autoridad competente
	T2	Explicabilidad ante afectados	Derecho exigible de las personas afectadas a recibir información significativa sobre el resultado
II. Responsabilidad proactiva	R1	Delimitación de roles	Obligaciones diferenciadas para proveedor, desplegador y usuario a lo largo de la cadena de valor
	R2	Mecanismos <i>ex ante</i>	Debida diligencia previa al despliegue con planes de contingencia y canales de reporte de incidentes
III. Mitigación de sesgos	S1	Evaluación previa	Evaluación obligatoria de sesgos y riesgos antes del despliegue
	S2	Seguimiento y participación	Monitoreo posterior al despliegue con participación de los grupos afectados
IV. Interacción	H1	Supervisión humana	Supervisión humana obligatoria y proporcional al riesgo
	H2	Veto vinculante	Autoridad efectiva del operador para anular el resultado, con protocolo de escalamiento definido

La codificación fue realizada íntegramente por el autor, dado que se trata de un estudio de autoría única, por lo que no fue posible estimar la concordancia entre codificadores. Para atenuar el riesgo de subjetividad que ello implica, se adoptaron tres salvaguardas. En primer lugar, cada nivel de la escala se definió mediante criterios explícitos y fijados antes de la codificación (Tabla 3), de modo que la valoración se apoyara en reglas y no en impresiones generales. En segundo lugar, cada puntuación se vinculó con las disposiciones concretas que la sustentan (Tabla 5), lo que permite a cualquier lector comprobar o rebatir la valoración asignada. En tercer lugar, la matriz completa y los fragmentos normativos codificados se depositaron en un repositorio abierto para facilitar su replicación independiente. Posteriormente, los fragmentos codificados se sometieron a un análisis temático [48] orientado a identificar patrones recurrentes, contradicciones y vacíos compartidos, que constituyeron el insumo directo para el diseño de los mecanismos del marco.

4.3 Demostración y estrategia de evaluación

Una vez construido, el artefacto se demostró mediante su aplicación a un escenario ilustrativo de servicios públicos de protección social, elegido porque combina en un mismo sistema resultados de bajo impacto y decisiones que afectan derechos, y porque la administración pública es uno de los ámbitos donde las exigencias de supervisión humana han mostrado mayores limitaciones [41]. Para la evaluación se adoptó, dentro del marco FEDS [49], una estrategia formativa, *ex ante* y artificial, apropiada para artefactos conceptuales en su primera iteración, que se concretó mediante los métodos descriptivos de argumentación informada y análisis de escenarios propuestos por Hevner et al. [16]. De esta manera, cada requisito de diseño se contrastó con los mecanismos del marco y con las evidencias que estos generan en el escenario, mientras que la evaluación naturalista en organizaciones reales se reserva para una iteración posterior. Dado que la investigación no involucró sujetos humanos ni datos personales, no requirió la aprobación de un comité de ética institucional. Aun así, conviene explicitar desde ahora que el análisis se limitó a documentos disponibles en español e inglés y a tres jurisdicciones predominantes, lo que omite perspectivas del Sur Global, y que existe un posible sesgo occidental en la selección de la teoría de base, que se procuró atenuar incorporando de manera deliberada perspectivas críticas en la discusión.

5. Resultados

Esta sección sigue el mismo orden del proceso de diseño y responde, de manera sucesiva, a cuatro preguntas. La primera es qué fallas presentan los marcos regulatorios vigentes (Sección 5.1), la segunda cómo responde el marco propuesto a esas fallas (Sección 5.2), la tercera qué aporta exactamente frente a lo existente (Sección 5.3) y la cuarta si funciona cuando se aplica a un caso concreto (Sección 5.4). De este modo, cada apartado se apoya en el anterior y prepara el siguiente, lo que permite seguir el razonamiento desde el diagnóstico hasta la aplicación.

5.1 Vacíos de los marcos regulatorios vigentes

El punto de partida consistió en examinar, con la matriz descrita en la Sección 4.2, hasta qué punto las normas de la Unión Europea, Estados Unidos y China contemplan los ocho mecanismos que el estudio considera necesarios para una IAG íntegra. Cada mecanismo recibió 0 puntos si la norma no lo prevé, 1 punto si lo prevé de forma parcial o solo como recomendación y 2 puntos si lo exige de manera obligatoria y verificable. La Tabla 4 presenta el resultado y la Tabla 5 recoge las disposiciones que sustentan cada puntuación, de modo que la valoración pueda ser replicada o discutida por otros investigadores.

Tabla 4. Puntuaciones de la matriz de evaluación por indicador y marco regulatorio

Indicador	UE	EE. UU.	China
T1 Documentación técnica	2	1	1
T2 Explicabilidad ante afectados	1	1	0
R1 Delimitación de roles	2	1	1
R2 Mecanismos <i>ex ante</i>	1	1	1
S1 Evaluación previa	2	1	1
S2 Seguimiento y participación	1	1	0
H1 Supervisión humana	2	1	1
H2 Veto vinculante	1	0	0
Total (0-16)	12	7	5

0 = ausencia, 1 = previsión parcial, 2 = previsión plena y vinculante.

Tabla 5. Disposiciones normativas que sustentan las puntuaciones de la Tabla 4

Indicador	UE (Reglamento 2024/1689)	EE. UU. (NIST AI RMF 1.0)	China (disposiciones 2021-2023)
T1 Documentación técnica	(2) Art. 11 y Anexo IV; art. 53 para modelos de uso general	(1) Documentación recomendada en las funciones MAP y MEASURE	(1) Registro de algoritmos y explicaciones sobre datos de entrenamiento a solicitud de la autoridad
T2 Explicabilidad ante afectados	(1) Art. 86, limitado a decisiones basadas en sistemas de alto riesgo del Anexo III; art. 50	(1) Características de sistemas explicables, interpretables y transparentes, con carácter recomendado	(0) Información general sobre el servicio, sin derecho a explicación del resultado individual
R1 Delimitación de roles	(2) Arts. 16, 25 y 26 (proveedor, cadena de valor y responsable del despliegue)	(1) GOVERN 2.1 y GOVERN 6 (roles y terceros, recomendados)	(1) Obligaciones concentradas en el proveedor del servicio, sin diferenciación a lo largo de la cadena
R2 Mecanismos <i>ex ante</i>	(1) Arts. 9, 20 y 73, limitados a alto riesgo y sin fondo de compensación	(1) MANAGE 2.4 y MANAGE 4 (recomendados)	(1) Evaluación de seguridad previa y mecanismos de quejas, orientados a la autoridad
S1 Evaluación previa	(2) Art. 10.2, letras f) y g), y art. 9	(1) MEASURE 2.11 (recomendado)	(1) Obligación general de prevenir la discriminación en el diseño y en los datos
S2 Seguimiento y participación	(1) Art. 72 y art. 27, sin participación vinculante de los afectados	(1) GOVERN 5.1 y MANAGE 4.1 (recomendados)	(0) Seguimiento orientado al control de contenidos, sin participación de los afectados
H1 Supervisión humana	(2) Art. 14 y art. 26.2	(1) GOVERN 3.2 (recomendado)	(1) Revisión de contenidos exigida al proveedor, orientada al control estatal
H2 Veto vinculante	(1) Art. 14.4, letras d) y e), sin protocolo de escalamiento definido	(0) Sin previsión de veto del operador sobre resultados individuales	(0) Sin previsión

Los resultados revelan tres perfiles regulatorios claramente diferenciados. El marco europeo alcanza la mayor cobertura, con 12 de 16 puntos, porque convierte en obligación la documentación técnica, el reparto de deberes entre quien desarrolla y quien utiliza el sistema, la evaluación previa al despliegue y la supervisión humana en los usos de alto riesgo. Sin embargo, pierde puntos precisamente donde la norma debería ceder capacidad de decisión a las personas, es decir, en el derecho de los afectados a recibir explicaciones, en su participación durante el seguimiento y en la facultad del operador para anular un resultado, lo que coincide con la crítica de Laux et al. [10]. El marco estadounidense, con 7 puntos, recomienda casi todos los mecanismos, pero no obliga a ninguno ni contempla el veto humano. El régimen chino, con 5 puntos, concentra su cobertura en el control estatal y carece de previsiones orientadas a los afectados, en consonancia con la prioridad que otorga a la estabilidad social [50]. La Figura 2 agrupa estos valores por eje y muestra que la interacción humano-IA es el eje con menor puntuación agregada entre los tres regímenes, con 5 de 12 puntos posibles, frente a 6 en transparencia y en sesgos y 7 en responsabilidad. Además, su indicador de veto vinculante (H2) es el menos cubierto de toda la matriz, con 1 de 6 puntos posibles, puesto que solo el marco europeo lo contempla y únicamente de forma parcial.

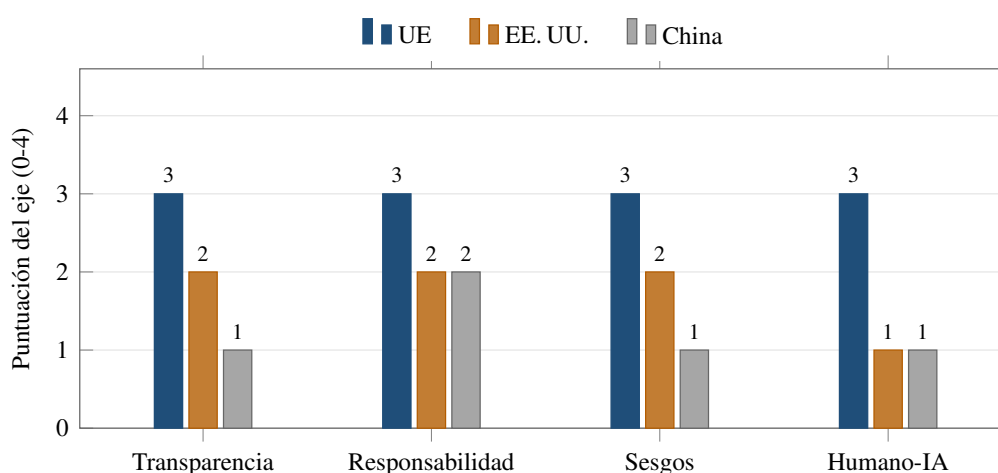


Figura 2. Perfil de cobertura normativa por eje en los tres marcos regulatorios analizados

Ahora bien, más allá de estas diferencias de intensidad, el análisis reveló cuatro carencias comunes a los tres regímenes. En primer lugar, la transparencia se dirige a la autoridad y no permite que un tercero reconstruya cómo se produjo un resultado concreto. En segundo lugar, la responsabilidad se asigna por tipo de actor, pero no se establece qué hacer si el modelo desarrolla comportamientos imprevistos. En tercer lugar, los sesgos se tratan como un problema técnico en el que las comunidades afectadas no intervienen y, por último, la supervisión humana se exige en términos genéricos, sin asegurar que el operador tenga autoridad, tiempo y formación para rechazar una decisión. Por tanto, estos resultados indican que las normas vigentes describen qué debe hacerse, pero no aseguran que pueda comprobarse que se hizo, y es precisamente este vacío de verificación el que da origen al marco que se presenta a continuación.

5.2 Marco de integridad como respuesta a los vacíos

En esta subsección se presenta el marco de integridad propuesto, cuyo propósito es dar respuesta a los cuatro vacíos identificados en el análisis anterior. Para ello, el marco transforma cada vacío en un eje de actuación y organiza los cuatro ejes alrededor de un núcleo común, la integridad centrada en el humano, sustentado a su vez en los tres pilares teóricos expuestos en la Sección 3, tal como se ilustra en la Figura 3. La lógica que articula el conjunto es sencilla, puesto que cada eje incorpora mecanismos concretos y cada mecanismo genera una evidencia que un tercero puede verificar, de manera que la integridad del sistema deja de ser una declaración de buenas intenciones y se convierte en una propiedad comprobable. Asimismo, las flechas discontinuas de la figura muestran que los ejes se retroalimentan entre sí, como ocurre cuando un caso escalado por un operador revela un sesgo que obliga a revisar los datos de entrenamiento.

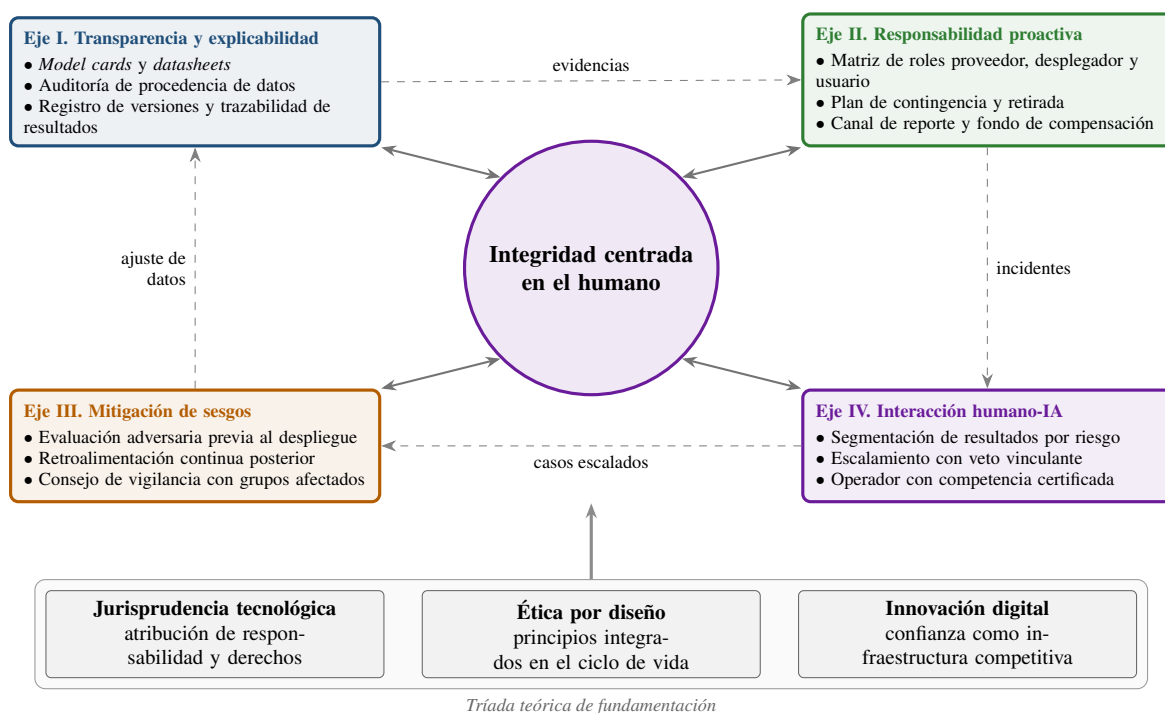


Figura 3. Marco de integridad ético-tecnológica para la IA generativa.

Para facilitar su lectura, cada eje se describe a continuación siguiendo el mismo recorrido, que parte del problema que atiende, continúa con lo que el marco propone y cierra con el sustento que respalda la propuesta.

5.2.1 Eje I: transparencia algorítmica y explicabilidad

El primer eje parte de una constatación, según la cual publicar información sobre un sistema no equivale a poder comprenderlo ni a exigir responsabilidades sobre él [38], y en los modelos de lenguaje esta dificultad se agrava porque generan texto a partir de patrones estadísticos, sin comprender su significado [5]. Frente a este problema, el eje sustituye la simple divulgación por la auditabilidad, entendida como la posibilidad de que un tercero reconstruya cómo se obtuvo un resultado. Para ello exige tres instrumentos, que son una ficha de modelo que describa su rendimiento, sus limitaciones y sus usos previstos [51], una ficha de datos que documente el origen y la composición de la información de entrenamiento [52] y un registro de versiones que vincule cada resultado con el modelo y la instrucción que lo produjeron. Esta explicabilidad por diseño se apoya en las técnicas de IA explicable [53] y responde a la obligación jurídica de ofrecer información significativa sobre las decisiones automatizadas que Kaminski [24] identifica en el derecho europeo.

5.2.2 Eje II: responsabilidad proactiva

El segundo eje se ocupa de un problema distinto. En la cadena de la IAG intervienen varios actores, entre ellos quien entrena el modelo, quien lo adapta y quien lo utiliza, y cuando se produce un daño la responsabilidad tiende a diluirse entre ellos. Además, estos modelos pueden desarrollar capacidades que nadie anticipó durante su entrenamiento [2]. Para evitarlo, el eje define las responsabilidades antes del despliegue y no después del daño. Siguiendo a Hacker et al. [17], los riesgos que nacen en el entrenamiento corresponden al proveedor del modelo y los que surgen en su adaptación o uso corresponden a quien lo despliega. A esta matriz de roles se suma un plan de contingencia, exigible cada vez que se libera un modelo, que especifica cómo retirarlo, por qué canal se reportan los daños y con qué fondo se compensa a los afectados. Esta propuesta se justifica porque la asignación temprana de responsabilidades a lo largo del ciclo de vida evita su dispersión [54], la decisión de cómo liberar un modelo constituye en sí misma una decisión de gobernanza [43] y las auditorías externas garantizan que la evaluación no quede en manos de quien se beneficia de ella [55, 19].

5.2.3 Eje III: mitigación de sesgos y derechos fundamentales

El tercer eje reconoce que ningún algoritmo es neutral, puesto que los datos con los que aprende reproducen desigualdades existentes [56, 57], y muchos estudios sobre sesgos ni siquiera definen qué daño pretenden medir [39]. Para enfrentar este problema, el eje combina medidas técnicas y sociales. En el plano técnico incorpora la evaluación adversaria, que consiste en someter el sistema a pruebas diseñadas para provocar respuestas dañinas antes de ponerlo en funcionamiento [58], así como la auditoría del origen de los datos y la recogida continua de reportes de los usuarios. En el plano social crea un consejo de vigilancia con representantes de los grupos afectados, encargado de decidir qué se considera un resultado injusto y cómo se mide. La dimensión social resulta indispensable, dado que riesgos como la desinformación o los estereotipos no se resuelven solo con ajustes técnicos [6], y la participación de los afectados solo es efectiva cuando les otorga capacidad real de decisión y no se limita a una consulta formal [40, 59].

5.2.4 Eje IV: interacción humano-IA

El cuarto eje aborda una idea que suele darse por supuesta, pues incluir a una persona en el proceso no garantiza que esta controle al sistema. Green [41] mostró que los supervisores humanos a menudo no pueden cumplir la función que se les asigna, y la literatura sobre el sesgo de automatización indica que las personas tienden a aceptar las recomendaciones de la máquina sin verificarlas, sobre todo cuando tienen exceso de trabajo [60, 42]. Para que la supervisión sea real, el eje establece un protocolo de escalamiento que clasifica cada resultado según su riesgo para las personas. Los resultados de bajo riesgo se entregan directamente y se revisa una muestra después, mientras que los de riesgo intermedio se entregan con una advertencia y se revisan con prioridad. Los de alto riesgo, así como las decisiones que por su naturaleza nunca pueden delegarse a la máquina, requieren la aprobación previa de un operador certificado cuyo veto es vinculante, es decir, que el sistema no puede ignorar. De este modo, la supervisión humana se concentra donde realmente importa y no se diluye entre miles de respuestas triviales. El Algoritmo 1 formaliza el procedimiento, en el que los umbrales θ_b y θ_a delimitan los niveles de riesgo bajo, intermedio y alto. La función EVALUARRIESGO puede implementarse mediante reglas explícitas o mediante un modelo de clasificación, y tanto sus criterios como los umbrales deben fijarse en cada contexto de uso, por lo que la Sección 5.4 ofrece un ejemplo orientativo de su configuración. Con ello, el protocolo materializa la idea de una IA centrada en el humano, que combina alta automatización con alto control humano [61], y sigue la lógica de la auditoría por capas propuesta por Mökander et al. [18].

Algoritmo 1 Protocolo de escalamiento humano segmentado por riesgo

Entrada: Resultado candidato y generado ante la instrucción x , contexto de uso u , umbrales $\theta_b < \theta_a$, tasa de muestreo ρ

Salida: Decisión final d y registro auditable ℓ

```

1:  $r \leftarrow \text{EVALUARRIESGO}(x, y, u)$  ▷ Clasificador de impacto sobre derechos
2: si  $u$  pertenece a una categoría de decisión no delegable o  $r \geq \theta_a$  entonces
3:    $v \leftarrow \text{VALIDARPOROPERADOR}(y, \text{justificación})$  ▷ Validación previa obligatoria
4:   si  $v = \text{veto}$  entonces
5:      $d \leftarrow \text{rechazo}$  ▷ El veto humano es vinculante
6:      $\text{NOTIFICARCONSEJO}(x, y, \text{motivo})$ 
7:   si no
8:      $d \leftarrow y$ 
9:   fin si
10: si no si  $\theta_b \leq r < \theta_a$  entonces
11:    $d \leftarrow y$  con advertencia al usuario y marcado para revisión prioritaria
12: si no
13:    $d \leftarrow y$ 
14:   Incluir  $(x, y)$  en la muestra de revisión posterior con probabilidad  $\rho$ 
15: fin si
16:  $\ell \leftarrow \text{REGISTRAR}(x, y, r, d, \text{operador}, t)$ 
17: devolver  $(d, \ell)$ 

```

La Figura 4 muestra el mismo protocolo desde la perspectiva de los actores involucrados. El recuadro superior recoge el recorrido de un resultado de alto riesgo, que pasa por el operador antes de llegar al usuario, y el inferior el de un resultado de bajo riesgo, que se entrega de inmediato y queda registrado para su revisión. La figura permite apreciar, además, que el protocolo conecta los cuatro ejes, ya que el registro alimenta la transparencia (Eje I), la notificación al consejo activa la responsabilidad y la revisión de sesgos

(Ejes II y III) y el veto hace efectiva la supervisión humana (Eje IV).

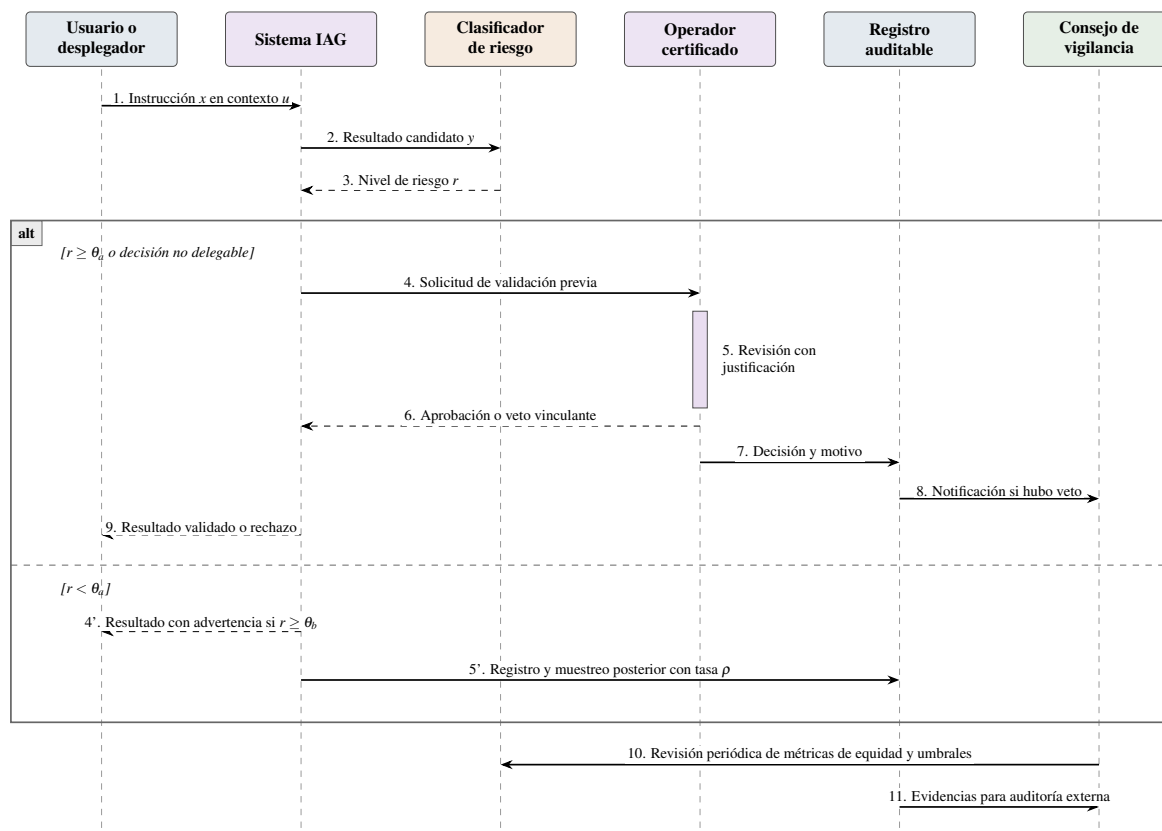


Figura 4. Diagrama de secuencia del protocolo de escalamiento humano.

De esta manera, lo expuesto pone de relieve que los cuatro ejes no funcionan como piezas aisladas, sino como un sistema en el que la evidencia generada por uno alimenta a los demás, una característica que resulta central para valorar la contribución del marco.

5.3 Contribución del marco frente a lo existente

Una vez descritos los ejes y su articulación, conviene hacer explícita la contribución del marco. La Tabla 6 lee de izquierda a derecha el recorrido completo del estudio, que va desde cada vacío detectado en las normas, pasa por el mecanismo que el marco propone para cubrirlo y termina en la evidencia concreta que permite comprobar su cumplimiento.

Tabla 6. Correspondencia entre vacíos identificados, mecanismos del marco y evidencias auditables

Vacío identificado	Eje	Mecanismo del marco	Evidencia verificable
Transparencia orientada a la autoridad y no a la trazabilidad de resultados	I	Explicabilidad por diseño con registro de versiones y auditoría de procedencia	Fichas de modelo y de datos, registro de versiones y traza reproducible de resultados
Responsabilidad sin mecanismos previos ante capacidades imprevistas	II	Matriz de roles y plan de contingencia asociado a cada decisión de liberación	Protocolo de retirada, canal de incidentes y fondo de compensación constituido
Mitigación de sesgos reducida a una evaluación técnica	III	Evaluación adversaria y consejo de vigilancia con grupos afectados	Informes de evaluación adversaria, actas del consejo y métricas de equidad validadas
Supervisión humana genérica y sin autoridad efectiva	IV	Protocolo de escalamiento segmentado por riesgo con veto vinculante	Registro de escalamientos, tasa de vetos y certificación de los operadores

Como se desprende de la Tabla 6, la contribución no consiste en añadir nuevos principios éticos a una lista ya extensa, sino en vincular cada vacío con un mecanismo y cada mecanismo con una prueba verificable, lo que permite a reguladores, auditores y desarrolladores comprobar el cumplimiento sin depender de declaraciones de buena fe.

5.4 Aplicación del marco a un caso concreto

Para comprobar que el marco puede llevarse a la práctica, se aplicó a un escenario ilustrativo. Una entidad pública de protección social utiliza un asistente de IAG, construido sobre un modelo de otra empresa, para dos tareas. La primera consiste en responder consultas generales sobre requisitos y plazos y la segunda en redactar borradores de resolución sobre solicitudes de subsidio, que luego firma un funcionario. El caso resulta exigente porque combina en un mismo sistema respuestas de bajo impacto y decisiones que afectan derechos sociales. En este escenario, el Eje II se aplica en primer lugar, porque define quién responde por qué. El proveedor del modelo se hace cargo de documentar el entrenamiento, el integrador tecnológico del ajuste y los filtros de seguridad, y la entidad de decidir qué decisiones no pueden delegarse y de designar a los operadores, mientras que un plan de contingencia prevé suspender la redacción de resoluciones si se detectan patrones discriminatorios. Con esa base, el Eje I exige que cada borrador quede vinculado a la versión del modelo y a la instrucción que lo originó, de modo que cualquier resolución cuestionada pueda reconstruirse. El Eje III, por su parte, somete el sistema antes de su puesta en marcha a pruebas con solicitantes simulados de distintos territorios, niveles de alfabetización y variedades lingüísticas, y durante su uso un consejo con representantes de beneficiarios revisa, por ejemplo, si las denegaciones propuestas se concentran en algún grupo. Por último, el Eje IV establece que toda resolución denegatoria requiere la aprobación previa del funcionario, que puede vetarla, que las consultas personales complejas se entregan con advertencia y que las consultas generales se revisan por muestreo, con una tasa que a título ilustrativo podría fijarse en el 5 %.

Para hacer operativa la función EVALUARRIESGO del Algoritmo 1, en el escenario se propone, a título orientativo, un clasificador basado en reglas que asigna a cada resultado una puntuación $r \in [0, 1]$ a partir de tres criterios, que son el tipo de tarea (consulta general o borrador de resolución), la presencia de datos personales del solicitante y el efecto potencial del resultado sobre el acceso a una prestación. Con umbrales ilustrativos de $\theta_b = 0,3$ y $\theta_a = 0,7$, las consultas generales sin datos personales quedan en el nivel bajo, las consultas personales complejas en el intermedio y los borradores de resolución en el alto, mientras que las resoluciones denegatorias se tratan como una categoría no delegable con independencia de su puntuación. La calibración definitiva de estos valores corresponde a la entidad, con la validación del consejo de vigilancia, y debe revisarse periódicamente a partir del registro de escalamientos, tal como prevé el paso 10 de la Figura 4.

Con este escenario como referencia, se verificó a continuación si el marco cumple los cinco requisitos definidos en la Sección 4.1. Como resume la Tabla 7, dos requisitos se satisfacen plenamente, la auditabilidad y la responsabilidad definida de antemano, y tres se satisfacen de forma condicionada, porque su cumplimiento depende de factores que el marco puede habilitar pero no garantizar por sí solo, como que el consejo sea realmente representativo, que los operadores no estén sobrecargados y que la organización disponga de recursos para documentar.

Tabla 7. Evaluación analítica del artefacto frente a los requisitos de diseño en el escenario de demostración

Requisito	Mecanismo del marco	Evidencia en el escenario	Valoración
RD1 Auditabilidad	Fichas de modelo y de datos, registro de versiones y trazas	Cada proyecto de resolución puede reconstruirse a partir de la versión y la instrucción que lo originó	Satisfecho
RD2 Responsabilidad <i>ex ante</i>	Matriz de roles y plan de contingencia	Obligaciones del proveedor, el integrador y la entidad definidas antes del despliegue, con protocolo de reversión	Satisfecho
RD3 Participación de afectados	Consejo de vigilancia con grupos afectados	Validación de métricas de denegación por grupo poblacional	Condicionado a la representatividad del consejo
RD4 Supervisión efectiva	Protocolo de escalamiento con veto vinculante	Resoluciones denegatorias validadas por un funcionario con autoridad para anularlas	Condicionado a la carga de trabajo del operador
RD5 Proporcionalidad	Segmentación por niveles de riesgo	Consultas generales supervisadas por muestreo y decisiones sobre derechos con validación previa	Condicionado a la capacidad de la organización

Estos resultados permiten concluir que el marco es aplicable en un contexto de alto impacto y cubre por diseño los requisitos de verificación y responsabilidad, mientras que la participación, la supervisión efectiva y la proporcionalidad exigen además condiciones organizacionales que deberán comprobarse en su implementación, cuestión que se retoma en la discusión.

6. Discusión

Los resultados confirman que la brecha entre principios y prácticas descrita por Mittelstadt [12] y por Morley et al. [14] persiste en los regímenes regulatorios más influyentes, incluso en el europeo, que obtuvo la mayor cobertura. Este hallazgo matiza la lectura según la cual la aprobación del Reglamento (UE) 2024/1689 habría resuelto el problema de la operacionalización, puesto que la matriz revela que sus fortalezas se concentran en obligaciones formales de documentación y evaluación, mientras que sus debilidades aparecen justamente en los indicadores que exigen transferir capacidad de decisión a las personas afectadas o a los operadores humanos. Por tanto, los resultados convergen con la crítica de Laux et al. [10] y la precisan, pues identifican los mecanismos concretos en los que la confiabilidad declarada no se traduce en salvaguardas exigibles.

En términos de la tipología de Gregor y Hevner [36], la contribución se ubica en la categoría de mejora, puesto que ofrece una solución nueva a un problema ya reconocido, y su valor reside tanto en el artefacto como en los principios de diseño que lo sustentan, los cuales pueden trasladarse a otros contextos. Desde esta perspectiva, el marco propuesto dialoga con los aportes de la auditoría de modelos de lenguaje. Mökander et al. [18] mostraron que las auditorías coordinadas en tres capas permiten identificar y gestionar riesgos, y el presente estudio extiende esa lógica al tiempo de operación, ya que el protocolo de escalamiento genera durante el uso las evidencias que después examinan los auditores. De igual manera, la incorporación de consejos de vigilancia responde a la advertencia de Birhane et al. [40] sobre el riesgo de una participación meramente simbólica, al vincular a dichos consejos con decisiones concretas como la validación de métricas y la revisión de umbrales. Por su parte, el diseño del protocolo toma en serio la evidencia de Green [41] y de Goddard et al. [60], dado que no presupone que la presencia humana baste, sino que concentra la supervisión allí donde el riesgo lo justifica y dota al operador de autoridad y competencia para ejercerla.

En cuanto al cuarto propósito específico, relativo a las implicaciones para la innovación digital, su tratamiento es exploratorio, puesto que no se tradujo en un requisito de diseño evaluable en esta iteración. Con esa salvedad, el marco puede operar como un mecanismo de prevención de riesgos reputacionales y legales que reduce la incertidumbre asociada a los proyectos de IAG. Siguiendo la perspectiva de la gobernanza responsable como capacidad organizacional [20], la disponibilidad de evidencias auditables permite a las organizaciones demostrar su diligencia ante usuarios, inversionistas y reguladores, lo que convierte la ética en una fuente de confianza competitiva. Simultáneamente, el marco aborda la tensión entre seguridad y soberanía tecnológica que Smuha [7] y Veale et al. [8] describen como fragmentación regulatoria, en la medida en que sus indicadores ofrecen un lenguaje común que permite a cada jurisdicción ajustar los umbrales de riesgo a su contexto sociopolítico sin renunciar a la interoperabilidad de los mercados digitales. La verificación empírica de estos efectos, por ejemplo mediante estudios de caso con organizaciones que adopten el marco, queda como una línea de trabajo futuro.

Conviene reconocer, además, una dimensión que el marco integra solo de manera parcial. Strubell et al. [62] mostraron que el entrenamiento de modelos de gran escala conlleva costos energéticos y ambientales considerables, y Cows et al. [63] advirtieron que el potencial de la IA para afrontar el cambio climático debe sopesarse con su propia huella. Aunque el marco exige documentar estos costos en las fichas de modelo, la incorporación plena de la sostenibilidad como quinto eje requiere investigación específica que excede el alcance del presente estudio. Por último, el estudio presenta limitaciones que delimitan su alcance. La evaluación realizada es formativa y artificial, de modo que demuestra la coherencia y la operabilidad del artefacto en un escenario ilustrativo, pero no su eficacia en organizaciones reales, que solo una evaluación naturalista podrá establecer [49]. Además, las puntuaciones de la matriz reflejan la codificación de textos normativos y no su aplicación efectiva, por lo que un régimen con alta cobertura formal podría mostrar un cumplimiento real inferior, y el análisis se circunscribe a tres jurisdicciones, lo que deja fuera las per-

spectivas del Sur Global. A ello se suma que la codificación fue realizada por un único investigador, sin posibilidad de estimar la concordancia entre codificadores, limitación que se procuró compensar con criterios explícitos, trazabilidad de cada puntuación y acceso abierto a la matriz, y que la comparación combina instrumentos vinculantes con un marco voluntario, cuya puntuación está acotada por su propia naturaleza, que el perfil de IA generativa del NIST [47] no se codificó de forma independiente y que las disposiciones chinas se analizaron mediante traducciones no oficiales, lo que puede introducir matices interpretativos. En relación con el Sur Global, el marco asume una capacidad institucional de auditoría que muchos países latinoamericanos todavía no poseen, lo que podría generar efectos de exclusión si no se acompaña de mecanismos de cooperación técnica internacional. Asimismo, cabe considerar la crítica de Hagendorff [13] y de Floridi [29] sobre el riesgo de reducir la deliberación moral a una lista de verificación, riesgo que el marco procura atenuar al exigir evidencias sustantivas y participación de los afectados, aunque no lo elimina por completo. Sin embargo, estas limitaciones no invalidan la propuesta, pero indican que la siguiente iteración del ciclo de diseño debe consistir en su evaluación en organizaciones concretas.

7. Conclusiones

Este estudio diseñó, demostró y evaluó analíticamente, desde un enfoque de ciencia del diseño, un marco de integridad ético-tecnológica para la IA generativa articulado en cuatro ejes, que comprenden la transparencia y la explicabilidad, la responsabilidad proactiva, la mitigación de sesgos y la interacción humano-IA. El análisis documental del Reglamento (UE) 2024/1689, del marco del NIST y del régimen chino evidenció que, pese a sus diferencias, los tres regímenes comparten vacíos en la explicabilidad ante los afectados, en la participación social y, sobre todo, en el veto humano vinculante. La principal contribución del trabajo consiste en vincular cada uno de estos vacíos con un mecanismo operativo y con una evidencia verificable por terceros, e incorporar un protocolo de escalamiento segmentado por riesgo que hace compatible la supervisión humana efectiva con la escala de operación de estos sistemas, de modo que la ética se integra como variable de diseño y no como un control posterior. La demostración en un escenario de protección social confirmó que el marco satisface los requisitos de auditabilidad y de responsabilidad *ex ante*, aunque la participación, la supervisión efectiva y la proporcionalidad dependen de condiciones organizacionales que deberán garantizarse en su implementación. Dado que la evaluación realizada fue de carácter analítico, el trabajo futuro se orientará a validar el marco en organizaciones reales, calibrar empíricamente el clasificador de riesgo y sus umbrales, verificar sus efectos sobre la competitividad y la confianza, extender la matriz a otras jurisdicciones, incluidas las latinoamericanas, e incorporar la sostenibilidad ambiental como dimensión adicional.

Contribución del autor. W.G. realizó la conceptualización, la metodología, la investigación, el análisis formal, la redacción del borrador original, la revisión y edición del texto y la visualización, además de revisar los resultados y aprobar la versión final del manuscrito.

Agradecimientos. A la Universidad Nacional Experimental de los Llanos Occidentales Ezequiel Zamora por el apoyo institucional brindado para el desarrollo de esta investigación.

Financiamiento. El estudio ha sido autofinanciado por el autor.

Declaración de consentimiento informado. No aplica, dado que el estudio no involucró participantes humanos.

Declaración de disponibilidad de datos. La matriz de codificación de los marcos regulatorios, los fragmentos normativos codificados, las disposiciones que sustentan cada puntuación y la documentación del escenario de demostración se encuentran disponibles a solicitud.

Conflictos de intereses. El autor declara no tener conflictos de intereses.

Declaración de uso de inteligencia artificial. Se emplearon herramientas de inteligencia artificial generativa exclusivamente para apoyar la edición del texto y la maquetación en $\text{\LaTeX}$. El análisis, la interpretación de los resultados y las conclusiones son responsabilidad intelectual exclusiva del autor, quien validó y aprobó la versión final.

Referencias

- [1] J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos, “Compute trends across three eras of machine learning,” in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–8. doi: 10.1109/IJCNN55064.2022.9891914.
- [2] D. Ganguli, D. Hernandez, L. Lovitt, A. Askell, Y. Bai, A. Chen *et al.*, “Predictability and surprise in large generative models,” in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1747–1764. doi: 10.1145/3531146.3533229.
- [3] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar *et al.*, ““So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy,” *International Journal of Information Management*, vol. 71, p. 102642, 2023. doi: 10.1016/j.ijinfomgt.2023.102642.
- [4] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023. doi: 10.1145/3571730.
- [5] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610–623. doi: 10.1145/3442188.3445922.
- [6] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor *et al.*, “Taxonomy of risks posed by language models,” in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 214–229. doi: 10.1145/3531146.3533088.
- [7] N. A. Smuha, “From a ‘race to AI’ to a ‘race to AI regulation’: Regulatory competition for artificial intelligence,” *Law, Innovation and Technology*, vol. 13, no. 1, pp. 57–84, 2021. doi: 10.1080/17579961.2021.1898300.
- [8] M. Veale, K. Matus, and R. Gorwa, “AI and global governance: Modalities, rationales, tensions,” *Annual Review of Law and Social Science*, vol. 19, pp. 255–275, 2023. doi: 10.1146/annurev-lawsocsci-020223-040749.
- [9] M. Veale and F. Zuiderveen Borgesius, “Demystifying the Draft EU Artificial Intelligence Act: Analysing the good, the bad, and the unclear elements of the proposed approach,” *Computer Law Review International*, vol. 22, no. 4, pp. 97–112, 2021. doi: 10.9785/cr-2021-220402.
- [10] J. Laux, S. Wachter, and B. Mittelstadt, “Trustworthy artificial intelligence and the European Union AI Act: On the conflation of trustworthiness and acceptability of risk,” *Regulation & Governance*, vol. 18, no. 1, pp. 3–32, 2024. doi: 10.1111/rego.12512.
- [11] A. Jobin, M. Ienca, and E. Vayena, “The global landscape of AI ethics guidelines,” *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019. doi: 10.1038/s42256-019-0088-2.
- [12] B. Mittelstadt, “Principles alone cannot guarantee ethical AI,” *Nature Machine Intelligence*, vol. 1, no. 11, pp. 501–507, 2019. doi: 10.1038/s42256-019-0114-4.
- [13] T. Hagendorff, “The ethics of AI ethics: An evaluation of guidelines,” *Minds and Machines*, vol. 30, no. 1, pp. 99–120, 2020. doi: 10.1007/s11023-020-09517-8.
- [14] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, “From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices,” *Science and Engineering Ethics*, vol. 26, no. 4, pp. 2141–2168, 2020. doi: 10.1007/s11948-019-00165-5.
- [15] L. Floridi and J. Cowls, “A unified framework of five principles for AI in society,” *Harvard Data Science Review*, vol. 1, no. 1, 2019. doi: 10.1162/99608f92.8cd550d1.
- [16] A. R. Hevner, S. T. March, J. Park, and S. Ram, “Design science in information systems research,” *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004. doi: 10.2307/25148625.

- [17] P. Hacker, A. Engel, and M. Mauer, "Regulating ChatGPT and other large generative AI models," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 1112–1123. doi: 10.1145/3593013.3594067.
- [18] J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi, "Auditing large language models: A three-layered approach," *AI and Ethics*, vol. 4, no. 4, pp. 1085–1115, 2024. doi: 10.1007/s43681-023-00289-2.
- [19] C. Novelli, M. Taddeo, and L. Floridi, "Accountability in artificial intelligence: What it is and how it works," *AI & Society*, vol. 39, no. 4, pp. 1871–1882, 2024. doi: 10.1007/s00146-023-01635-y.
- [20] E. Papagiannidis, P. Mikalef, and K. Conboy, "Responsible artificial intelligence governance: A review and research framework," *The Journal of Strategic Information Systems*, vol. 34, no. 2, p. 101885, 2025. doi: 10.1016/j.jsis.2024.101885.
- [21] A. Matthias, "The responsibility gap: Ascribing responsibility for the actions of learning automata," *Ethics and Information Technology*, vol. 6, no. 3, pp. 175–183, 2004. doi: 10.1007/s10676-004-3422-1.
- [22] F. Santoni de Sio and G. Mecacci, "Four responsibility gaps with artificial intelligence: Why they matter and how to address them," *Philosophy & Technology*, vol. 34, no. 4, pp. 1057–1084, 2021. doi: 10.1007/s13347-021-00450-x.
- [23] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation," *International Data Privacy Law*, vol. 7, no. 2, pp. 76–99, 2017. doi: 10.1093/idpl/ixp005.
- [24] M. E. Kaminski, "The right to explanation, explained," *Berkeley Technology Law Journal*, vol. 34, no. 1, pp. 189–218, 2019. doi: 10.15779/Z38TD9N83H.
- [25] J. J. Bryson, "Patience is not a virtue: The design of intelligent systems and systems of ethics," *Ethics and Information Technology*, vol. 20, no. 1, pp. 15–26, 2018. doi: 10.1007/s10676-018-9448-6.
- [26] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, and L. Floridi, "Artificial intelligence and the 'good society': The US, EU, and UK approach," *Science and Engineering Ethics*, vol. 24, no. 2, pp. 505–528, 2018. doi: 10.1007/s11948-017-9901-7.
- [27] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, "AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018. doi: 10.1007/s11023-018-9482-5.
- [28] D. Schiff, J. Borenstein, J. Biddle, and K. Laas, "AI ethics in the public, private, and NGO sectors: A review of a global document collection," *IEEE Transactions on Technology and Society*, vol. 2, no. 1, pp. 31–42, 2021. doi: 10.1109/TTS.2021.3052127.
- [29] L. Floridi, "Translating principles into practices of digital ethics: Five risks of being unethical," *Philosophy & Technology*, vol. 32, no. 2, pp. 185–193, 2019. doi: 10.1007/s13347-019-00354-x.
- [30] A. Kasirzadeh and I. Gabriel, "In conversation with artificial intelligence: Aligning language models with human values," *Philosophy & Technology*, vol. 36, no. 2, p. 27, 2023. doi: 10.1007/s13347-023-00606-x.
- [31] S. Zuboff, "Big other: Surveillance capitalism and the prospects of an information civilization," *Journal of Information Technology*, vol. 30, no. 1, pp. 75–89, 2015. doi: 10.1057/jit.2015.5.
- [32] A. Birhane, "Algorithmic injustice: A relational ethics approach," *Patterns*, vol. 2, no. 2, p. 100205, 2021. doi: 10.1016/j.patter.2021.100205.
- [33] T. Birkstedt, M. Minkkinen, A. Tandon, and M. Mäntymäki, "AI governance: Themes, knowledge gaps and future agendas," *Internet Research*, vol. 33, no. 7, pp. 133–167, 2023. doi: 10.1108/INTR-01-2022-0042.

- [34] M. Whittaker, “The steep cost of capture,” *Interactions*, vol. 28, no. 6, pp. 50–55, 2021. doi: 10.1145/3488666.
- [35] M. Taddeo, T. McCutcheon, and L. Floridi, “Trusting artificial intelligence in cybersecurity is a double-edged sword,” *Nature Machine Intelligence*, vol. 1, no. 12, pp. 557–560, 2019. doi: 10.1038/s42256-019-0109-1.
- [36] S. Gregor and A. R. Hevner, “Positioning and presenting design science research for maximum impact,” *MIS Quarterly*, vol. 37, no. 2, pp. 337–355, 2013. doi: 10.25300/MISQ/2013/37.2.01.
- [37] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A design science research methodology for information systems research,” *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007. doi: 10.2753/MIS0742-1222240302.
- [38] M. Ananny and K. Crawford, “Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability,” *New Media & Society*, vol. 20, no. 3, pp. 973–989, 2018. doi: 10.1177/1461444816676645.
- [39] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, “Language (technology) is power: A critical survey of “bias” in NLP,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5454–5476. doi: 10.18653/v1/2020.acl-main.485.
- [40] A. Birhane, W. Isaac, V. Prabhakaran, M. Díaz, M. C. Elish, I. Gabriel, and S. Mohamed, “Power to the people? Opportunities and challenges for participatory AI,” in *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO ’22)*, 2022, pp. 1–8. doi: 10.1145/3551624.3555290.
- [41] B. Green, “The flaws of policies requiring human oversight of government algorithms,” *Computer Law & Security Review*, vol. 45, p. 105681, 2022. doi: 10.1016/j.clsr.2022.105681.
- [42] R. Parasuraman and D. H. Manzey, “Complacency and bias in human use of automation: An attentional integration,” *Human Factors*, vol. 52, no. 3, pp. 381–410, 2010. doi: 10.1177/0018720810376055.
- [43] I. Solaiman, “The gradient of generative AI release: Methods and considerations,” in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 111–122. doi: 10.1145/3593013.3593981.
- [44] G. A. Bowen, “Document analysis as a qualitative research method,” *Qualitative Research Journal*, vol. 9, no. 2, pp. 27–40, 2009. doi: 10.3316/QRJ0902027.
- [45] Parlamento Europeo y Consejo de la Unión Europea, “Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (Reglamento de Inteligencia Artificial),” *Diario Oficial de la Unión Europea*, Serie L, 12 de julio de 2024, 2024. [Online]. Available: <http://data.europa.eu/eli/reg/2024/1689/oj>
- [46] National Institute of Standards and Technology, “Artificial intelligence risk management framework (AI RMF 1.0),” National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep. NIST AI 100-1, 2023. doi: 10.6028/NIST.AI.100-1.
- [47] —, “Artificial intelligence risk management framework: Generative artificial intelligence profile,” National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep. NIST AI 600-1, 2024. doi: 10.6028/NIST.AI.600-1.
- [48] V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006. doi: 10.1191/1478088706qp063oa.
- [49] J. Venable, J. Pries-Heje, and R. Baskerville, “FEDS: A framework for evaluation in design science research,” *European Journal of Information Systems*, vol. 25, no. 1, pp. 77–89, 2016. doi: 10.1057/ejis.2014.36.

- [50] H. Roberts, J. Cows, J. Morley, M. Taddeo, V. Wang, and L. Floridi, "The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation," *AI & Society*, vol. 36, no. 1, pp. 59–77, 2021. doi: 10.1007/s00146-020-00992-2.
- [51] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 220–229. doi: 10.1145/3287560.3287596.
- [52] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021. doi: 10.1145/3458723.
- [53] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020. doi: 10.1016/j.inffus.2019.12.012.
- [54] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 33–44. doi: 10.1145/3351095.3372873.
- [55] I. D. Raji, P. Xu, C. Honigsberg, and D. Ho, "Outsider oversight: Designing a third party audit ecosystem for AI governance," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, pp. 557–571. doi: 10.1145/3514094.3534181.
- [56] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021. doi: 10.1145/3457607.
- [57] A. Birhane and V. U. Prabhu, "Large image datasets: A pyrrhic win for computer vision?" in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1536–1546. doi: 10.1109/WACV48630.2021.00158.
- [58] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving, "Red teaming language models with language models," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 3419–3448. doi: 10.18653/v1/2022.emnlp-main.225.
- [59] J. Metcalf, E. Moss, E. A. Watkins, R. Singh, and M. C. Elish, "Algorithmic impact assessments and accountability: The co-construction of impacts," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 735–746. doi: 10.1145/3442188.3445935.
- [60] K. Goddard, A. Roudsari, and J. C. Wyatt, "Automation bias: A systematic review of frequency, effect mediators, and mitigators," *Journal of the American Medical Informatics Association*, vol. 19, no. 1, pp. 121–127, 2012. doi: 10.1136/amiajnl-2011-000089.
- [61] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *International Journal of Human-Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020. doi: 10.1080/10447318.2020.1741118.
- [62] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for modern deep learning research," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 13 693–13 696. doi: 10.1609/aaai.v34i09.7123.
- [63] J. Cows, A. Tsamados, M. Taddeo, and L. Floridi, "The AI gambit: Leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations," *AI & Society*, vol. 38, no. 1, pp. 283–307, 2023. doi: 10.1007/s00146-021-01294-x.