Atribución-NoComercial-CompartirIgual 4.0 Internacional (CC BY-NC-SA 4.0)

Alberto Davis Ortiz[1]
Jorge Antonio Breceda Pérez[2]

https://doi.org/10.20983/anuariodcispp.2025.11

Fecha de recepción: 11 de febrero 2025
Fecha de aprobación: 02 de julio 2025

LA EXPLICABILIDAD DE LA IA (XAI) COMO IMPERATIVO CATEGÓRICO PARA LA SEGURIDAD JURÍDICA Y LA INTEGRIDAD ÉTICA

AI explainability (XAI) as a categorical imperative for legal certainty and ethical integrity

RESUMEN

La creciente incorporación de sistemas de inteligencia artificial opacos en decisiones públicas y privadas plantea un problema estructural para los derechos humanos: la imposibilidad de conocer las razones detrás de decisiones automatizadas erosiona la dignidad humana, vulnera el debido proceso y debilita la seguridad jurídica. El objetivo del artículo es demostrar que la explicabilidad (xai) constituye un imperativo categórico para garantizar la legitimidad ética y jurídica de la ia en contextos de alto impacto. Metodológicamente, se desarrolla un enfoque interdisciplinario que articula un análisis filosófico-normativo (dignidad kantiana e injusticia epistémica), revisión jurídico-comparada de jurisprudencia crítica (State v. Loomis, syri, Deliveroo) y un estudio técnico de herramientas de explicabilidad (lime, shap, contrafactuales) como instrumentos de auditoría y control. Los resultados evidencian que la opacidad algorítmica genera tres efectos sistémicos: cosificación del individuo, al ser reducido a un vector de datos; afectación estructural al derecho de defensa, al impedir impugnar la lógica de decisión; y discriminación indirecta derivada de variables proxy invisibles. Asimismo, los casos comparados muestran que la ausencia de explicabilidad permite que el secreto comercial prevalezca sobre garantías procesales, mientras que los modelos transparentes fortalecen el equilibrio entre eficacia administrativa y derechos fundamentales. El estudio concluye que la ia solo puede ser compatible con un orden constitucional democrático si incorpora mecanismos robustos de explicabilidad orientados no solo a describir el funcionamiento técnico, sino a justificar normativamente las decisiones.

1 Profesor investigador del Departamento de Ingeniería Eléctrica de la Universidad Autónoma de Ciudad Juárez (uacj). https://orcid.org/0000-0001-6840-5563
2 Profesor investigador adscrito al Departamento de Ciencias Jurídicas de la Universidad Autónoma de Ciudad Juárez (uacj). https://orcid.org/0000-0001-5280-6936

La xai emerge así como una condición estructural de legitimidad y el fundamento indispensable para la responsabilidad y confianza pública en sistemas algorítmicos.

Palabras clave: debido proceso; discriminación indirecta; explicabilidad algorítmica; responsabilidad tecnológica.

ABSTRACT

The increasing incorporation of opaque artificial intelligence systems in public and private decisions poses a structural problem for human rights: the inability to know the reasons behind automated decisions erodes human dignity, violates due process, and weakens legal certainty. The aim of this article is to demonstrate that explainability (xai) constitutes a categorical imperative to guarantee the ethical and legal legitimacy of ai in high-impact contexts. Methodologically, an interdisciplinary approach is developed that articulates philosophical-normative analysis (Kantian dignity and epistemic injustice), a comparative legal review of critical jurisprudence (State v. Loomis, syri, Deliveroo), and a technical study of explainability tools (lime, shap, counterfactuals) as instruments of auditing and control. The results show that algorithmic opacity generates three systemic effects: the objectification of the individual by reducing them to a data vector; structural impact on the right to a defense by preventing challenges to the decision-making logic; and indirect discrimination stemming from invisible proxy variables. Furthermore, the compared cases show that the lack of explainability allows trade secrecy to prevail over procedural guarantees, while transparent models strengthen the balance between administrative efficiency and fundamental rights. The study concludes that ai can only be compatible with a democratic constitutional order if it incorporates robust explainability mechanisms aimed not only at describing the technical operation, but also at normatively justifying the decisions. xai thus emerges as a structural condition of legitimacy and as the indispensable foundation for public accountability and trust in algorithmic systems.

Keywords: algorithmic explainability; due process; indirect discrimination; technological accountability.

LA EXPLICABILIDAD

DE LA IA (XAI) COMO IMPERATIVO CATEGÓRICO PARA LA SEGURIDAD JURÍDICA Y LA INTEGRIDAD ÉTICA

1. INTRODUCCIÓN

La integración de la inteligencia artificial (*IA*) en los estratos fundamentales de la toma de decisiones sociales, legales y médicas ha precipitado una crisis de legitimidad sin precedentes en la historia de la técnica. A diferencia de las revoluciones industriales precedentes, que automatizaron la fuerza física, la revolución algorítmica automatiza el juicio cognitivo y normativo. Sin embargo, la prevalencia de modelos de aprendizaje profundo (*Deep Learning*) y redes neuronales complejas ha dado lugar al fenómeno de la "caja negra" (*black box*): sistemas cuya construcción interna es tan opaca o compleja que sus procesos de inferencia resultan inescrutables, incluso para sus propios creadores (London, 2019).

Esta opacidad difiere radicalmente de la ignorancia técnica tradicional. No se trata simplemente de que el código sea privado (secreto comercial) o difícil de leer para un lego. Se trata de una barrera epistémica fundamental: en las redes neuronales profundas, la relación entre la entrada (*input*) y la salida (*output*) no sigue reglas lógicas lineales programadas por un humano, sino que emerge de millones de ajustes de parámetros autodidactas. Esto crea una paradoja regulatoria: tenemos sistemas que funcionan con una eficacia estadística asombrosa, pero cuyas razones para actuar son ininteligibles, rompiendo la cadena de causalidad necesaria para la atribución de responsabilidad jurídica.

Este informe técnico, elaborado desde una perspectiva multidisciplinaria que fusiona la tecno-ética, la filosofía del derecho y la regulación comparada, sostiene la tesis de que la Explicabilidad de la Inteligencia Artificial (*XAI*) no es meramente una funcionalidad técnica deseable o un "lujo" de ingeniería. Por el contrario, la *XAI* constituye un requisito indispensable para la preservación de la seguridad jurídica y la integridad ética.

La tensión dialéctica actual se debate entre dos argumentos polares: el "Argumento del Humano Terrible" (*Awful Human Argument*), que justifica la automatización opaca basándose en la falibilidad y los sesgos cognitivos humanos, y el "Argumento de Mejor Juntos" (*Better Together Argument*), que aboga por una simbiosis donde la máquina aumente la capacidad humana (Binns *et al.*, 2018). Sin embargo, esta investigación demuestra que sin una explicabilidad robusta, la colaboración es ilusoria y la automatización se convierte en una forma de dominación arbitraria.

La tensión dialéctica actual es crítica. Por un lado, el "Argumento del Humano Terrible" sugiere que, dado que los jueces humanos tienen hambre, celos, prejuicios y días malos, deberíamos preferir la consistencia fría de la máquina, incluso si es opaca. Por otro lado, el "Argumento de Me-

jor Juntos" propone una simbiosis donde la *IA* detecta patrones que el humano ignora, y el humano aporta el contexto moral que la *IA* no comprende. Sin embargo, esta investigación demuestra que sin una explicabilidad robusta (*XAI*), la colaboración del "Mejor Juntos" es una ficción, ya que el humano no colabora con la máquina, pues simplemente se subordina a ella por falta de argumentos para contradecirla.

A través del análisis de jurisprudencia internacional crítica —como *State v. Loomis* en Estados Unidos, *SyRI* en los Países Bajos y el caso Deliveroo en Italia— y la evaluación de marcos normativos emergentes, como la *AI* Act de la Unión Europea y los estándares *IEEE* P7000, este documento desglosa cómo la opacidad algorítmica erosiona el debido proceso, facilita la injusticia epistémica y crea "zonas de deformación moral", donde la responsabilidad se diluye. La transición de modelos de "caja negra" a sistemas de "caja de cristal" (*glass-box*) o, al menos, justificables, se presenta como el desafío regulatorio definitorio de nuestra era.

Para fundamentar esta tesis, el presente informe se estructura en tres ejes. Primero, examinamos la dimensión ontológica y ética (secciones 2 y 4), analizando cómo la opacidad vulnera la dignidad humana y facilita la discriminación. Segundo, abordamos la dimensión legal y de seguridad jurídica (secciones 3 y 5), contrastando la jurisprudencia crítica, como Loomis y *SyRI*, para ilustrar la tensión entre la eficiencia y el debido proceso. Finalmente, exploramos la dimensión bioética y de confianza pública (secciones 6 y 7), proponiendo la transición de la mera "explicación técnica" a la "justificación normativa" como el nuevo estándar de oro para la inteligencia artificial democrática.

2. EL IMPERATIVO ÉTICO

El primer eje crítico de esta investigación no es técnico, sino ontológico. Refiere al estatus del ser humano frente a la máquina que lo evalúa, clasifica o sentencia. La ética de la *IA*, a menudo trivializada en listas de principios vagos, debe anclarse en una defensa robusta de la dignidad humana contra la cosificación algorítmica.

La filosofía moral kantiana establece una distinción categórica que resulta vital para la tecno-ética moderna:

> En el reino de los fines, todo tiene un precio o una dignidad. Lo que tiene un precio puede ser sustituido por algo equivalente; lo que, en cambio, se halla por encima de todo precio y, por tanto, no admite nada equivalente, tiene dignidad (*Würde*). (Mäki-Kuutti, Raisamo y Vakkuri, 2021, p. 3).

La adopción de sistemas de *IA* opacos en la gestión de asuntos humanos amenaza con colapsar esta distinción. Cuando un algoritmo reduce a un individuo a un

UACJ - Anuario de derecho, comercio internacional, seguridad y políticas públicas Año 4, Número 4, 2025,
No. de Reserva: 04-2022-101412422200-102

vector de datos (*data points*) y toma decisiones que alteran su vida, basándose en correlaciones estadísticas invisibles, el sujeto es tratado como un medio para un fin (eficiencia administrativa, maximización de beneficios, seguridad predictiva) y no como un fin en sí mismo. La falta de explicabilidad priva al individuo de la capacidad de comprender las razones de su tratamiento, eliminando la posibilidad de consentimiento racional o disenso fundamentado, características esenciales de la agencia moral (Mäki-Kuutti *et al.*, 2021).

Recientes marcos regulatorios, como la Ley de *IA* de la *UE* (Reglamento 2024/1689), han comenzado a incorporar la dignidad como un valor aspiracional y un derecho positivo amenazado por prácticas de *IA*, como la manipulación subliminal, el *nudging* (empujoncito) coercitivo y la vigilancia biométrica (Keeling, 2024). La dignidad exige que el ser humano no sea objeto de un proceso computacional opaco, sino sujeto de un procedimiento inteligible. Sin *XAI*, la interacción humano-máquina se degrada a una relación de objeto a sujeto, donde la máquina "actúa" y el humano "padece".

Más allá de la dignidad intrínseca, la opacidad algorítmica genera una forma específica y perniciosa de daño ético: la injusticia epistémica. Este concepto, originado en la filosofía social de Miranda Fricker y adaptado al contexto tecnológico, describe el daño que sufren los individuos en su capacidad como sujetos de conocimiento (*knowers*) (Alami *et al.*, 2024); por ello, es importante mencionar:

Primero, se produce cuando se otorga un exceso de credibilidad a la salida del algoritmo (percibido como objetivo, matemático y neutral), mientras se devalúa sistemáticamente el testimonio o la experiencia vivida del sujeto afectado (Alami *et al.*, 2024); por ello, es dable considerar:

a) Mecanismo: un "déficit de credibilidad" injustificado se asigna al humano frente a la máquina.
b) Ejemplo: un sistema de detección de fraude en prestaciones sociales marca a un ciudadano como sospechoso. Aunque el ciudadano presente pruebas de su honestidad, la "autoridad" de la señal algorítmica prevalece, y el ciudadano es tratado con sospecha *a priori*. La opacidad del sistema impide al ciudadano desafiar la base epistémica de la acusación, silenciando efectivamente su testimonio (Alami *et al.*, 2024).

Segundo, se produce cuando la opacidad del sistema impide que los afectados comprendan siquiera la naturaleza de su desventaja, privándoles de los recursos conceptuales necesarios para dar sentido a su experiencia (Viljoen y Wenger, 2024).

a) Mecanismo: se crea una "laguna epistémica" donde la discriminación o el error son invisibles e ininteligibles.
b) Contexto: si una *IA* de contratación rechaza sistemáticamente a mujeres cualificadas debido a una correlación oculta (por ejemplo, brechas en el historial laboral debido a la maternidad, codificadas como "falta de compromiso"), las candidatas pueden percibir su rechazo como un fracaso personal. Carecen de la explicación causal (el sesgo del modelo) que les permitiría interpretar su situación como una injusticia estructural en lugar de una insuficiencia individual (Alami *et al.*, 2024).

La respuesta a este imperativo ético se ha cristalizado en estándares internacionales que elevan la transparencia a principio rector. El marco "Ethically Aligned Design" del IEEE (Estándar P7000) posiciona la transparencia no como una característica técnica, sino como un pilar para garantizar los derechos humanos y el bienestar (*IEEE* Standards Association, 2019). El *IEEE* introduce la noción de que la base de una decisión de un sistema autónomo debe ser siempre "descubrible" (*discoverable*). Esto implica una trazabilidad desde el diseño hasta el despliegue, asegurando que los operadores sean responsables y rindan cuentas (Verity*AI*, 2024).

Paralelamente, la Recomendación sobre la Ética de la Inteligencia Artificial de la Unesco (adoptada por 193 Estados miembros) establece la "Transparencia y Explicabilidad" como uno de sus diez principios centrales (Unesco, 2021). La Unesco subraya que la explicabilidad es fundamental para garantizar que los sistemas de *IA* no socaven los derechos humanos y las libertades fundamentales. Es notable que la Unesco vincula explícitamente la falta de transparencia con la imposibilidad de auditar violaciones de derechos humanos, haciendo de la *XAI* una condición *sine qua non* para la gobernanza democrática (Unesco, 2022).

3. SEGURIDAD JURÍDICA

Es el principio que garantiza que el derecho sea conocido, previsible y aplicable con certeza. La introducción de algoritmos predictivos opacos en el sistema judicial y administrativo plantea un desafío directo a este principio, creando lo que académicos como Danielle Citron han denominado la necesidad de un "debido proceso tecnológico" (Citron, 2019).

Si en la sección anterior establecimos que la opacidad algorítmica lesiona la dignidad del sujeto moral, en esta sección analizaremos cómo erosiona las garantías del sujeto legal. La dignidad humana, traducida al lenguaje procesal, se manifiesta en el derecho a la defensa y al contradictorio. Sin embargo, la seguridad jurídica —la certeza sobre cómo y por qué se aplica la ley— enfrenta una amenaza existencial

UACJ - Anuario de derecho, comercio internacional, seguridad y políticas públicas Año 4, Número 4, 2025,
No. de Reserva: 04-2022-101412422200-102

cuando el "juez" o el "evaluador" administrativo opera bajo una lógica inescrutable.

El debido proceso exige que cuando el Estado priva a una persona de vida, libertad o propiedad, debe proporcionar una notificación adecuada y una oportunidad significativa para ser escuchado. ¿Puede haber una "oportunidad significativa" de defensa si la evidencia en contra del acusado es una puntuación de riesgo generada por un algoritmo secreto?

La investigación destaca que la *IA* viola el debido proceso procedimental si el gobierno se niega a revelar las razones de una decisión automatizada (fianza, beneficios sociales, estatus migratorio) (Citron, 2019). Los sistemas de "caja negra" convierten el proceso legal en un ejercicio kafkiano, donde el acusado lucha contra una lógica invisible. La "autenticidad de la evidencia" y la "credibilidad de los testigos" (en este caso, el testigo digital) se ven socavadas si no se puede interrogar al algoritmo sobre sus tasas de error, sus datos de entrenamiento y sus sesgos latentes (Siegel y Klein, 2024).

El caso *State v. Loomis* (2016) ante la Corte Suprema de Wisconsin es paradigmático de la tensión entre propiedad intelectual corporativa y derechos constitucionales (Harvard Law Review, 2017). Lo anterior se puede sintetizar en la tabla 1.

Este fallo tuvo consecuencias que se extienden mucho más allá de Wisconsin. Al validar el uso de proxies de riesgo opacos bajo el escudo del secreto comercial, la Corte sentó un precedente peligroso: la propiedad intelectual corporativa fue

Tabla 1.

Categoría	Descripción técnico-jurídica
Hechos	Eric Loomis fue condenado utilizando como uno de los insumos una puntuación de riesgo de reincidencia generada por el *software* propietario COMPAS. El acusado alegó violación al debido proceso, al no poder acceder al algoritmo para verificar su precisión, su validez científica o los posibles sesgos derivados de su entrenamiento, dada la naturaleza cerrada y comercialmente protegida del sistema (Washington, 2019).
Fallo	La Corte Suprema de Wisconsin resolvió que el uso de COMPAS no vulnera el debido proceso si se cumplen condiciones mínimas: 1) que la puntuación no sea el elemento determinante único en la sentencia; 2) que exista una advertencia escrita para los jueces sobre las limitaciones metodológicas del sistema; y 3) que se reconozca su carácter evaluador de riesgos grupales y no individuales (Supreme Court of Wisconsin, 2016).
Crítica y análisis doctrinal	La decisión ha sido fuertemente cuestionada por legitimar la opacidad algorítmica. La Corte privilegió el secreto comercial por encima del derecho de defensa y del control epistémico del imputado sobre la evidencia utilizada en su contra. La utilidad de las "advertencias" ha sido puesta en duda ante el sesgo de anclaje: la etiqueta "alto riesgo" influye en la decisión judicial, incluso si se advierte sobre sus limitaciones. El caso revela una afectación a la seguridad jurídica, donde la eficiencia administrativa se antepone a la transparencia y a los estándares de debido proceso (Harvard Law Review, 2017).

Fuente: elaboración propia.

priorizada por encima del derecho fundamental del acusado a confrontar la evidencia en su contra. Loomis ilustra una "seguridad jurídica debilitada", donde el sistema prioriza la eficiencia administrativa sobre la transparencia radical, dejando al individuo indefenso ante una caja negra que define su libertad.

En contraste directo, el caso SyRI (2020) en los Países Bajos representa el triunfo de la transparencia como precondición de legalidad (Rachovitsa, 2022), que se puede leer en la tabla 2.

A diferencia de Loomis, el caso SyRI establece una doctrina de "transparencia como precondición de legalidad". El tribunal reconoció que la asimetría de información era tan severa que rompía el equilibrio justo entre el Estado y el ciudadano. La implicación profunda de este fallo es que la eficacia de un algoritmo para detectar fraude es irrelevante si su funcionamiento no puede ser auditado públicamente. El SyRI nos enseña que en un Estado de derecho no basta con que la decisión sea correcta, pues el camino para llegar a ella debe ser visible. Dicho análisis se visualiza en la tabla 3.

Por otra parte, la Unión Europea ha intentado codificar estas lecciones en la Ley

Tabla 2.

Categoría	Descripción técnico-jurídica
Hechos	El gobierno de los Países Bajos implementó el Sistema de Indicación de Riesgo (SyRI) para detectar fraudes en prestaciones de seguridad social mediante el cruce masivo y automatizado de datos provenientes de múltiples agencias públicas. Diversas organizaciones de la sociedad civil interpusieron una demanda al considerar que el sistema generaba un esquema de vigilancia desproporcionado y opaco hacia la ciudadanía.
Fallo	El Tribunal de Distrito de La Haya declaró que la normativa que sustentaba el SyRI violaba el artículo 8 del Convenio Europeo de Derechos Humanos, al afectar ilegítimamente el derecho a la vida privada. El tribunal sostuvo que la arquitectura legal que permitía el funcionamiento del SyRI no cumplía los estándares mínimos de protección frente a la injerencia estatal en datos personales (Rachovitsa, 2022).
Fundamento jurídico y argumentativo	El tribunal no proscribió el uso estatal de algoritmos, pero enfatizó que el SyRI carecía de transparencia verificable, impidiendo a las personas conocer las razones por las cuales eran clasificadas como riesgos. Esta asimetría informacional rompía el "equilibrio justo" entre el interés público en combatir el fraude y los derechos fundamentales de los individuos. La opacidad tecnológica anulaba la posibilidad de defensa efectiva, configurando una violación estructural del Estado de derecho (Van Bekkum y Borgesius, 2021).
Implicación doctrinal	El caso establece que la transparencia algorítmica constituye un requisito constitucional para la legalidad de sistemas de vigilancia administrativa intensiva. El SyRI se convierte así en un precedente europeo clave que afirma que la legitimidad del uso estatal de IA requiere inteligibilidad y auditabilidad de la lógica de decisión.

Fuente: elaboración propia.

UACJ - Anuario de derecho, comercio internacional, seguridad y políticas públicas Año 4, Número 4, 2025,
No. de Reserva: 04-2022-101412422200-102

Tabla 3.

Criterio	State v. Loomis (2016, EE. UU.)	SYRI (2020, Países Bajos)
Contexto y hechos relevantes	Uso judicial de COMPAS, un sistema propietario de evaluación de riesgo penal. El acusado no pudo conocer la lógica del algoritmo ni cuestionar su validez científica.	Implementación estatal del sistema SYRI para detectar fraudes a partir del cruce masivo de datos. Ciudadanía afectada sin posibilidad de conocer criterios de clasificación.
Problema jurídico central	Tensión entre el secreto comercial y el derecho al debido proceso y defensa efectiva.	Tensión entre la vigilancia algorítmica estatal y el derecho a la vida privada (art. 8 CEDH).
Decisión judicial	La Corte validó el uso de COMPAS con advertencias y límites, sin exigir transparencia plena.	El tribunal declaró ilegal el sistema por falta de transparencia y proporcionalidad.
Estándar aplicado	Aceptación de herramientas opacas, siempre que no sean el factor determinante y se advierta al juzgador sobre sus limitaciones.	Exigencia de "transparencia verificable" como condición para la legalidad del tratamiento automatizado de datos.
Acceso del afectado a la lógica algorítmica	Negado: prevalece el secreto comercial del desarrollador.	Requerido: la persona debe poder comprender por qué fue clasificada como riesgo.
Impacto en el debido proceso/ defensa	Debilitado: el acusado no puede refutar la evidencia algorítmica ni conocer sus tasas de error.	Fortalecido: el tribunal considera indispensable la posibilidad de defensa frente a decisiones automatizadas.
Peso otorgado a los derechos fundamentales	Derechos procesales subordinados a la eficiencia judicial y a la propiedad intelectual.	Prevalencia del derecho a la vida privada y al Estado de derecho sobre la eficiencia administrativa.
Implicación doctrinal	Precedente que legitima la opacidad algorítmica y normaliza la evidencia no auditable.	Precedente europeo que establece la transparencia algorítmica como requisito constitucional.
Modelo de gobernanza tecnológica que fomenta	"Caja negra tolerada" con advertencias; confianza en el operador humano sin inteligibilidad.	"Caja de cristal" obligatoria; algoritmos auditables, inteligibles y sujetos a control judicial.

Fuente: elaboración propia.

de Inteligencia Artificial. El Reglamento clasifica ciertos usos (como la evaluación de riesgo en justicia penal o acceso a servicios públicos) como de "Alto Riesgo", imponiendo obligaciones estrictas de transparencia (Henriksen, 2024).

El artículo 14 introduce el requisito de "Vigilancia Humana" (*Human Oversight*). Exige que los sistemas se diseñen de tal manera que los humanos puedan "entender adecuadamente las capacidades y limitaciones" del sistema y monitorear su operación para detectar anomalías (European Union, 2024).

Sin embargo, el análisis crítico sugiere que este artículo podría crear una "tram-

pa legal". Al transferir la responsabilidad de la supervisión al operador humano (el "humano en el bucle"), sin garantizar que la IA sea técnicamente explicable, se corre el riesgo de crear una figura de paja que valida decisiones opacas sin entenderlas realmente (EU AI Act, 2024). La regulación exige transparencia, pero su efectividad dependerá de si se implementa como mera documentación técnica o verdadera inteligibilidad para el operador.

4. JUSTICIA ALGORÍTMICA Y LA AUDITORÍA DE LA DISCRIMINACIÓN

La justicia algorítmica busca asegurar la equidad (*fairness*) y la no discriminación en las decisiones automatizadas. La opacidad de las cajas negras es el mayor obstáculo para este fin, ya que permite que los sesgos se oculten bajo capas de complejidad matemática.

Existe la creencia errónea de que un algoritmo es justo si no utiliza variables protegidas (raza, género) explícitamente. Sin embargo, a través de variables proxy (código postal, historial de navegación), los modelos de *Deep Learning* pueden reconstruir estas categorías y discriminar con alta precisión (Barocas & Selbst, 2017). Sin XAI, estos sesgos permanecen invisibles hasta que producen daño masivo.

Para ello, el caso del Tribunal de Bolonia contra la plataforma de reparto Deliveroo (2020) es ilustrativo de cómo la falta de contexto y explicabilidad conduce a la discriminación laboral (Aloisi y De Stefano, 2021), el cual se puede explicar de la siguiente manera (tabla 4).

El caso Deliveroo cristaliza el concepto de injusticia hermenéutica discutido anteriormente. El algoritmo "Frank" no estaba programado para discriminar, simplemente estaba programado para ser ciego al contexto humano (enfermedad, huelga). Esta "ceguera" técnica se tradujo en una discriminación real. Este caso demuestra que la neutralidad algorítmica es un mito: un sistema que no puede explicar ni entender las razones detrás de una acción humana (contexto) inevitablemente penalizará los comportamientos legítimos que se desvían de la norma estadística.

Para combatir estos sesgos, las herramientas técnicas de XAI deben reconfigurarse como instrumentos de auditoría moral y legal (tabla 5).

5. RESPONSABILIDAD MORAL Y LEGAL

La atribución de responsabilidad (*liability*) es uno de los problemas más espinosos en la regulación de la IA. ¿Quién es responsable cuando una IA autónoma causa daño? La falta de explicabilidad rompe la cadena causal necesaria para establecer negligencia o dolo.

La antropóloga Madeleine Elish acuñó el término "Moral Crumple Zone" (Zona de Deformación Moral) para describir un fenómeno inquietante en sistemas comple-

UACJ - Anuario de derecho, comercio internacional, seguridad y políticas públicas Año 4, Número 4, 2025,
No. de Reserva: 04-2022-101412422200-102

Tabla 4.

Categoría	Descripción técnico-jurídica
Hechos	El Tribunal de Bolonia examinó el sistema algorítmico "Frank", implementado por Deliveroo para evaluar la "fiabilidad" de los repartidores mediante puntuaciones que condicionaban su acceso a turnos prioritarios. El algoritmo reducía la calificación de quienes cancelaban turnos con menos de veinticuatro horas de anticipación, sin diferenciar las causas, lo que impactaba directamente en sus condiciones de trabajo (Aloisi y De Stefano, 2021).
Funcionamiento del algoritmo	"Frank" operaba bajo un modelo de análisis uniforme que penalizaba toda cancelación sin tomar en cuenta factores contextuales. Esto implicaba una evaluación automatizada puramente correlacional: cualquier cancelación equivalía a menor fiabilidad, sin distinguir entre decisiones voluntarias y causas protegidas legalmente (enfermedad, huelga, fuerza mayor).
Problema jurídico identificado	La "ceguera contextual" del algoritmo —incapacidad para reconocer razones legítimas— generó un trato desigual injustificado. Al no incorporar justificaciones relevantes, el sistema producía efectos discriminatorios contra repartidores que ejercían sus derechos fundamentales, pese a que los trataba formalmente igual (Clifford Chance, 2021).
Fallo	El Tribunal de Bolonia determinó que el sistema constituía discriminación indirecta, al penalizar desproporcionadamente a trabajadores protegidos por causas legalmente reconocidas. Ordenó que la puntuación algorítmica dejara de considerar las cancelaciones sin distinguir su motivación, al violar derechos laborales básicos.
Lección para XAI	El caso evidencia que la "neutralidad algorítmica" es un mito: un sistema que no contextualiza reproduce injusticias estructurales. Una arquitectura explicable y auditable habría permitido detectar *ex ante*, un criterio de evaluación incompatible con los derechos laborales fundamentales. La XAI se convierte así en herramienta de prevención de discriminación y garantía de debido proceso algorítmico.

Fuente: elaboración propia.

Tabla 5.

Herramienta XAI	Descripción técnica	Aplicación jurídico-normativa	Función en la justicia algorítmica
LIME (Local Interpretable Model-agnostic Explanations)	Genera modelos lineales locales que aproximan el comportamiento del modelo complejo en torno a una instancia concreta. Explica cómo variables específicas influyeron en un resultado individual (Liu, 2024).	Permite identificar la razón concreta por la cual una persona recibió una decisión adversa (rechazo de crédito, clasificación de riesgo, etcétera). Facilita la trazabilidad individual de la decisión.	Habilita la contestabilidad individual, al permitir determinar si la decisión se basó en variables legítimas o en proxies discriminatorios (por ejemplo, código postal como sustituto de raza o nivel socioeconómico).
SHAP (SHapley Additive exPlanations)	Calcula la contribución marginal de cada variable utilizando valores de Shapley. Proporciona explicaciones globales y locales del modelo (IEEE Computer Society, 2024).	Es fundamental para auditorías regulatorias: identifica qué atributos del modelo tienen mayor peso y detecta la influencia de atributos prohibidos, sesgos estructurales o correlaciones espurias.	Permite detectar discriminación indirecta y sesgos sistémicos, garantizando transparencia estructural. Es crucial para la supervisión estatal y el cumplimiento normativo en sistemas de alto riesgo.
Análisis contrafactual	Ofrece escenarios alternativos al responder: "¿qué tendría que cambiar para obtener un resultado diferente?". Evalúa rutas mínimas de cambio para modificar la salida del modelo (Deep Science Research, 2024).	Permite al individuo entender cómo podría obtener un resultado favorable en el futuro y demuestra si el sistema ofrece vías de mejora accesibles, no discriminatorias ni imposibles de cumplir.	Restaura la agencia del usuario al brindarle rutas claras de acción. Es esencial para el derecho a impugnar, para el debido proceso tecnológico y para la equidad en decisiones algorítmicas.

Fuente: elaboración propia.

jos: la responsabilidad por los fallos tiende a recaer en los operadores humanos más cercanos a la acción (médicos, pilotos, conductores), incluso cuando el sistema les negó el control efectivo o la información necesaria para evitar el error (Elish, 2016).

Al igual que la zona de deformación de un coche absorbe el impacto físico para proteger a los pasajeros, el operador humano absorbe el impacto legal y moral para proteger la integridad del sistema tecnológico y a sus creadores corporativos. Sin XAI, el operador no puede defenderse argumentando que el sistema le indujo a error, ya que la "caja negra" no revela su lógica defectuosa. El humano se convierte en el "fusible" que salta ante la catástrofe (Akitra, 2024).

En este sentido, la regulación actual (como el Art. 14 del AI Act) confía excesivamente en el "Humano en el Bucle" (HITL) como salvaguarda (European Union, 2024). Sin embargo, la evidencia psicológica sobre la interacción humano-máquina sugiere que esta supervisión es a menudo defectuosa, debido a dos sesgos cognitivos reforzados por la falta de explicabilidad (tabla 6).

Tabla 6.

Concepto/ Categoría	**Descripción técnico-jurídica y cognitiva**	**Efecto sobre la supervisión humana**	**Implicación para la regulación y la XAI**
Supuesto regulatorio: HITL (Human in the Loop)	El Art. 14 del AI Act exige supervisión humana significativa como salvaguarda para sistemas de alto riesgo. Se asume que un operador humano puede detectar errores, corregir sesgos y contradecir decisiones algorítmicas (European Union, 2024).	La supervisión depende de la capacidad del operador para comprender las salidas del sistema. Si la IA es opaca, el operador carece de herramientas para cuestionar su razonamiento.	La efectividad del HITL queda condicionada a la existencia de explicabilidad robusta; sin ella, la supervisión es meramente formal y no sustantiva.
Sesgo de Automatización (Automation Bias)	Tendencia cognitiva a confiar en la recomendación de la IA, incluso cuando existe información humana contradictoria y correcta. Surge de la percepción de que la máquina es objetiva, analítica y menos falible (Brown y Weiß, 2024).	El operador acepta la recomendación por defecto, omite verificar inconsistencias y reduce su juicio crítico.	La falta de explicabilidad potencia este sesgo: sin razones explícitas de la IA, el humano no tiene base epistémica para contradecirla.
Complacencia de la Automatización (Automation Complacency)	Fenómeno en el que la fiabilidad percibida del sistema provoca una disminución de la vigilancia humana. El operador actúa como un "sellador de goma", aprobando mecánicamente decisiones automatizadas (Liu *et al.*, 2023).	Aumenta la probabilidad de errores no detectados y decisiones injustas, pues el humano deja de monitorear activamente la lógica del sistema.	La XAI mitiga la complacencia al obligar al operador a interpretar, evaluar y contrastar explicaciones, incrementando el escrutinio cognitivo.
Caso ilustrativo de fallo HITL en medicina	Un algoritmo de seguros recomendó dar de alta a una paciente usuaria de silla de ruedas a un hogar en un quinto piso sin ascensor. La IA consideró únicamente la estabilidad médica, ignorando el contexto físico y social. El operador aprobó el alta sin cuestionar la recomendación (Parasuraman y Manzey, 2010).	El operador, afectado por complacencia y ausencia de explicabilidad contextual, no identificó la laguna lógica. La decisión produjo un daño grave.	Ilustra que la supervisión humana sin explicabilidad es una ficción regulatoria: la IA requiere ofrecer razones comprensibles y contextualizadas para habilitar una supervisión efectiva.

Fuente: elaboración propia.

UACJ - Anuario de derecho, comercio internacional, seguridad y políticas públicas Año 4, Número 4, 2025,
No. de Reserva: 04-2022-101412422200-102

Para salir de este atolladero, la investigación propone transitar hacia un ecosistema de responsabilidad compartida habilitado por la XAI, con el siguiente paradigma de responsabilidad (tabla 7).

6. CONFIANZA PÚBLICA Y EPISTEMOLOGÍA

La confianza pública es el capital social necesario para el despliegue de cualquier tecnología. En el caso de la IA, esta confianza está fracturada por el fenómeno de la Aversión Algorítmica: la tendencia de las personas a perder la confianza en un algoritmo más rápidamente que en un humano tras ver un error, incluso si el algoritmo es estadísticamente superior en general (Dietvorst, Simmons y Massey, 2015).

Estudios empíricos demuestran que permitir a los usuarios modificar ligeramente las predicciones de la IA o entender su margen de error reduce la aversión. La transparencia sobre la precisión y el funcionamiento interno (XAI) ayuda a los usuarios a "calibrar" su confianza: saber cuándo confiar y cuándo ser escéptico en lugar de una desconfianza o una fe ciegas (He, Yang y Chen, 2024). Un aporte crucial de la filosofía de la IA es la distinción entre estos dos conceptos, a menudo confundidos (tabla 8).

Tabla 7.

Actor	Ámbito de responsabilidad	Contenido técnico-jurídico de la obligación	Justificación normativa y funcional
Desarrollador	Responsabilidad en el diseño y arquitectura del sistema	No se limita al código fuente; incluye la obligación positiva de integrar mecanismos de explicabilidad. El desarrollador debe garantizar que el sistema pueda comunicar su lógica de decisión, sus límites operativos y los factores que influyen en cada resultado.	Sin explicabilidad incorporada desde el diseño (ex ante) es imposible reconstruir la cadena causal que fundamenta la responsabilidad. La transparencia técnica es condición para la auditabilidad y para evitar la "zona de deformación moral".
Desplegador (instituciones públicas, empresas, entidades usuarias)	Responsabilidad organizacional y de implementación	Debe asegurar que los operadores comprendan las explicaciones proporcionadas por el sistema mediante capacitación en alfabetización algorítmica. Asimismo, debe crear un entorno en el que contradecir al sistema no genere sanciones formales o informales.	El contexto institucional determina si la supervisión humana es real o simbólica. Sin incentivos para el juicio crítico, el hitl se convierte en mera legitimación pasiva del sistema.
Operador humano (jueces, personal médico, funcionarios, analistas)	Responsabilidad en la supervisión y decisión final	Ejercer un escrutinio crítico basado en las explicaciones del sistema. Esta responsabilidad solo puede activarse cuando la ia brinda razones comprensibles y contrastables.	El operador es el último eslabón del proceso decisional: su responsabilidad depende de que cuente con insumos inteligibles. La xai restituye su agencia y evita que funcione como un "sellador de goma".

Fuente: elaboración propia.

Tabla 8.

Concepto	Descripción técnico-filosófica	Relevancia para el usuario o ciudadano	Implicación jurídico-normativa y para la XAI
Calibración de la confianza mediante XAI	Estudios empíricos demuestran que cuando los usuarios pueden ajustar ligeramente las predicciones de la IA o visualizar el margen de error, disminuye la aversión algorítmica. La transparencia operativa permite calibrar la confianza: saber cuándo confiar y cuándo ser escéptico en lugar de adoptar fe ciega o desconfianza absoluta (He *et al.*, 2024).	Empodera al usuario al brindarle criterios para evaluar la fiabilidad de la IA y reduce la sensación de arbitrariedad tecnológica.	Refuerza el Principio de Autonomía Decisional y el derecho a comprender la evidencia algorítmica que afecta derechos, garantizando un debido proceso tecnológico sustantivo.
Explicación (Explanation)	Explicación causal y descriptiva: responde a: "¿cómo llegó la IA a este resultado?". Revela la secuencia factual de activaciones, pesos o reglas internas del sistema (Stanford Encyclopedia of Philosophy, 2015).	Es útil para ingenieros, pero suele ser irrelevante para el ciudadano, pues describe mecanismos internos sin ofrecer razones normativamente comprensibles.	No satisface por sí sola los estándares constitucionales de transparencia ni garantiza contestabilidad ni control democrático.
Justificación (Justification)	Responde a: "¿es esta una buena razón para tomar esta decisión en nuestra sociedad?". Identifica razones normativas aceptables que legitiman la decisión con base en principios públicos (Hadfield *et al.*, 2019).	Permite al ciudadano evaluar si la decisión es legítima, justa y coherente con normas éticas, laborales, administrativas o constitucionales.	Es el estándar necesario para decisiones públicas o de alto impacto: sin justificación, una decisión explicable puede seguir siendo incompatible con derechos humanos.
*Insigh*t de tercer orden: hacia una "IA justificable"	La meta de la tecnoética no es solo una IA explicable técnicamente, sino una IA capaz de generar razones normativamente válidas. Un sistema puede ser explicable (mostrar que discriminó por raza) y aun así ser inaceptable moral y jurídicamente. La XAI es el medio técnico para auditar la validez de las justificaciones (Hadfield *et al.*, 2019).	Permite comprender si la decisión afecta derechos o reproduce discriminación estructural, independientemente de su corrección técnica.	Sienta las bases para un modelo de gobernanza, donde la IA no solo sea transparente, sino compatible con el Estado constitucional de derecho, la igualdad y la dignidad humana.

Fuente: elaboración propia.

7. CONCLUSIONES

La investigación exhaustiva de los ejes éticos, legales y técnicos conduce a una conclusión unívoca: la Explicabilidad de la Inteligencia Artificial (XAI) no es un componente opcional, sino el cimiento sobre el cual debe construirse cualquier despliegue legítimo de IA en una sociedad democrática. A lo largo de este análisis, hemos demostrado tres proposiciones fundamentales que desafían el determinismo tecnológico actual:

UACJ - Anuario de derecho, comercio internacional, seguridad y políticas públicas Año 4, Número 4, 2025,
No. de Reserva: 04-2022-101412422200-102

I. Hemos establecido, siguiendo la ética kantiana, que la opacidad algorítmica instrumentaliza al ser humano, reduciéndolo a un medio (un vector de datos) y negándole la agencia necesaria para comprender y consentir su tratamiento. Esto genera una forma de violencia epistémica donde el sujeto es evaluado, pero no escuchado.
II. El contraste entre los casos Loomis y SyRI revela que la "Caja Negra" es incompatible con el Estado de derecho. No existe un verdadero derecho a la defensa si la evidencia acusatoria (el perfil de riesgo) es un secreto comercial inescrutable. La justicia requiere no solo un resultado, sino un procedimiento inteligible.
III. El análisis del caso Deliveroo y los sesgos ocultos demuestra que un algoritmo "ciego" al contexto es inherentemente discriminatorio. Sin XAI, estos sesgos operan en la sombra, protegidos por la complejidad matemática, perpetuando injusticias estructurales bajo una apariencia de objetividad técnica.

Ahora bien, para transitar de una "IA de Caja Negra" a una "IA Justificable", se proponen las siguientes medidas imperativas:

I. Es necesario ir más allá de los principios éticos voluntarios (soft law) hacia marcos vinculantes como la ai Act, pero con una salvedad crítica: la transparencia no puede ser meramente técnica (código abierto), debe ser epistémica. La regulación debe exigir que el sistema provea razones comprensibles para el afectado ("justificación normativa") y no solo mapas de neuronas ("explicación causal").
II. Siguiendo el precedente de syri, cualquier sistema utilizado por el Estado para asignar recursos o penalizar ciudadanos debe ser auditable por diseño. Recomendamos la implementación obligatoria de herramientas como shap y análisis contrafactuales en la fase de pruebas para detectar discriminación indirecta antes del despliegue.
III. Debemos desmantelar la "Zona de Deformación Moral" que culpa injustamente al operador humano. La legislación debe establecer que, si un sistema no ofrece explicaciones claras y contextuales, el desarrollador comparte la responsabilidad por los errores de juicio del operador, mitigando así el sesgo de automatización y la complacencia.

Sin duda, la transición hacia una IA explicable no es un lujo académico ni un freno a la innovación: es el prerrequisito para mantener la compatibilidad entre el progreso tecnológico y la civilización democrática. Los casos examinados nos muestran dos futuros posibles: el camino de Loomis, donde la eficiencia corporativa silenció los derechos civiles, o el cami-

no de SyRI, donde la transparencia actuó como salvaguarda de la libertad.

Por último, la pregunta fundamental que enfrenta nuestra generación legal y técnica no es si podemos permitirnos el "costo" computacional de la XAI, sino si podemos permitirnos el costo social y moral de su ausencia. Una IA que no puede explicar sus razones es, en última instancia, una forma de dominación arbitraria. Solo una IA justificable merece nuestra confianza y nuestro consentimiento.

REFERENCIAS

Akitra. (2024). *Accountability and liability in agentic AI systems.* https://akitra.com/accountability-and-liability-in-agentic-ai-systems/

Alami, H., Lehoux, P., Shaw, J., Fortin, J.-P., Fleet, R., Ag Ahmed, M. A., & Denis, J.-L. (2024). Epistemic injustice in generative AI. *AAAI/ACM Conference on AI, Ethics, and Society.* https://ojs.aaai.org/index.php/AIES/article/view/31671/33838

Aloisi, A., & De Stefano, V. (2021). Frankly, my rider, I don't give a damn. *La rivista il Mulino.* https://www.rivistailmulino.it/a/frankly-my-rider-i-don-t-give-a-damn-1

Barocas, S., & Selbst, A. D. (2017). Big data's disparate impact. *Colorado Technology Law Journal*, *15*(4). http://ctlj.colorado.edu/wp-content/uploads/2021/02/17.1_4-Washington_3.18.19.pdf

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). AI, algorithms, and awful humans. *Fordham Law Review*, *87*(6), 2147-2161. https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=6079&context=flr

Brown, A., & Weiß, M. (2024). Bias in the loop: How humans evaluate AI-generated suggestions. *arXiv.* https://arxiv.org/html/2509.08514v1

Business & Human Rights Resource Centre. (2021). *Court rules Deliveroo used "discriminatory" algorithm.* https://www.business-humanrights.org/en/latest-news/court-rules-deliveroo-used-discriminatory-algorithm/

Citron, D. K. (2019). Artificial intelligence and procedural due process. *University of Pennsylvania Journal of Constitutional Law*, *22*(1). https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=1901&context=jcl

Clifford Chance. (2021). The Italian courts lead the way on explainable AI. *Talking Tech.* https://www.cliffordchance.com/insights/resources/blogs/talking-tech/en/articles/2021/06/the-italian-courts-lead-the-way-on-explainable-ai.html

Deep Science Research. (2024). Explainable artificial intelligence (XAI) as a foundation for trustworthy artificial intelligence. *Deep Science Research.* https://deepscienceresearch.com/dsr/catalog/book/10/chapter/74

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114-126.

Elish, M. C. (2016). *Moral crumple zones: Cautionary tales in human-robot interaction.* Data & Society Research Institute.

EU AI Act. (2024). Key issue 4: Human oversight. https://www.euaiact.com/key-issue/4

European Union. (2024). Article 14: Human oversight. *EU Artificial Intelligence Act.* https://artificialintelligenceact.eu/article/14/

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, *46*(3), 205-211.

Hadfield, G. K., Bozdag, E., Law, A., & Neilson, S. (2019). Explanation and justification: AI decision-making, law, and the rights of citizens. *Schwartz Reisman Institute.* https://srinstitute.utoronto.ca/news/hadfield-justifiable-ai

Harvard Law Review. (2017). *State v. Loomis*, 130, 1530-1539. https://harvardlawreview.org/print/vol-130/state-v-loomis/

Hatherley, J. J. (2024). Healthy mistrust: Medical black box algorithms, epistemic authority, and preemptionism. *Cambridge Quarterly of Healthcare Ethics.* https://www.cambridge.org/core/journals/cambridge-quarterly-of-healthcare-ethics/article/healthy-mistrust-medical-black-box-algorithms-epistemic-authority-and-preemptionism/38018A52AF77F8C120DC815A4EE6AD52

He, Y., Yang, Q., & Chen, S. (2024). Algorithm appreciation or aversion: The effects of accuracy disclosure on users' reliance on algorithmic suggestions. *Behaviour & Information Technology.* https://www.tandfonline.com/doi/full/10.1080/0144929X.2025.2535732

Henriksen, A. (2024). High-risk AI transparency? On qualified transparency mandates for oversight bodies under the EU AI Act. *Technology and Regulation.* https://techreg.org/article/view/19876

IEEE Computer Society. (2024). AI's role in ethical decision-making: Fostering fairness in critical systems with explainable AI (XAI). *IEEE Computer Society.* https://www.computer.org/publications/tech-news/community-voices/explainable-ai

IEEE Standards Association. (2019). *Ethically aligned design.* http://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

Keeling, G. (2024). Why dignity is a troubling concept for AI ethics. *AI & Society.* https://pmc.ncbi.nlm.nih.gov/articles/PMC11963102/

Lauridsen, K. M., & Bjørnsen, H. N. (2024). Epistemic authority and medical AI: Epistemological differences and challenges in medical practice. *Medicine, Health Care and Philosophy.* https://www.researchgate.net/publication/397176188

Liu, S. (2024). Does explainable AI have moral value? *arXiv.* https://arxiv.org/html/2311.14687

Liu, M., Grunde-McLaughlin, M., Goel, A., & Brummette, M. (2023). Automation complacency: Navigating the ethical challenges of AI in healthcare. *Columbia University School of Professional Studies.* https://sps.columbia.edu/news/automation-complacency-navigating-ethical-challenges-ai-healthcare

London, A. J. (2019). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, *45*(12), 820-826. https://pubmed.ncbi.nlm.nih.gov/33737318/

Mäki-Kuutti, I., Raisamo, R., & Vakkuri, V. (2021). Philosophical foundations for digital ethics and AI ethics: A dignitarian approach. *Frontiers in Computer Science*, *3*. https://pmc.ncbi.nlm.nih.gov/articles/PMC7909376/

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381-410. https://pubmed.ncbi.nlm.nih.gov/21077562/

Rachovitsa, M. (2022). Human rights implications of the use of AI in the digital welfare state: Lessons learned from the Dutch SyRI case. *Human Rights Law Review, 22*(2). https://academic.oup.com/hrlr/article/22/2/ngac010/6568079

Raknes, S., & Bakken, T. H. (2023). Informed consent to AI-based decisions in healthcare: Must patients understand the AI's output? *Oslo Law Review*, *11*(1), 82-99. https://www.scup.com/doi/full/10.18261/olr.11.1.7

Siegel, M. D., & Klein, D. (2024). Deepfakes in the courtroom: Problems and solutions. *Illinois State Bar Association.* https://www.isba.org/sections/ai/newsletter/2025/03/deepfakesinthecourtroomproblemsandsolutions

Stanford Encyclopedia of Philosophy. (2015). Reasons for action: Justification vs. explanation. https://plato.stanford.edu/archives/sum2015/entries/reasons-just-vs-expl/

Supreme Court of Wisconsin. (2016). State v. Loomis, 881 N.W.2d 749. https://courts.ca.gov/sites/default/files/courts/default/2024-12/btb24-2l-3.pdf

United Nations Educational, Scientific and Cultural Organization (Unesco). (2021). *Recommendation on the ethics of artificial intelligence.* https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

-----. (2022). Unesco's input in reply to the OHCHR report on the Human Rights Council Resolution 47/23. https://www.ohchr.org/sites/default/files/2022-03/UNESCO.pdf

Van Bekkum, M., & Borgesius, F. Z. (2021). Digital welfare fraud detection and the Dutch SyRI judgment. *Computer Law & Security Review*, *42.* https://www.iapp.org/news/a/digital-welfare-fraud-detection-and-the-dutch-syri-judgment

VerityAI. (2024). IEEE ethically aligned design: Engineering ethics into AI systems. https://verityai.co/blog/ieee-ethically-aligned-design-guide

Viljoen, S., & Wenger, A. (2024). Algorithmic profiling as a source of hermeneutical injustice. *Philosophy & Technology*, *38*(1). https://pmc.ncbi.nlm.nih.gov/articles/PMC11741985/

Washington, A. L. (2019). How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. *Colorado Technology Law Journal*, *17*(1). http://ctlj.colorado.edu/wp-content/uploads/2021/02/17.1_4-Washington_3.18.19.pdf