Principios de inteligencia artificial explicable (XAI), en el diseño ético de soluciones de ciberdefensa en las Fuerzas Militares colombianas.

John Deivy Diaz Narvaez
Principios de inteligencia artificial explicable (XAI), en el diseño ético de soluciones de ciberdefensa en las Fuerzas militares colombianas, Escuela Superior de Guerra "General Rafael Reyes Prieto", Colombia. Ingeniero de telecomunicaciones, Especialista en seguridad Informática, Msc Ciberseguridad y Ciberdefensa, Msc Diseño y Gestión de Proyectos Tecnológicos. Alumno CIM2025, Escuela Superior de Guerra "General Rafael Reyes Prieto", Colombia. https://orcid.org/0009-0009-7786-4484.
Contacto: jhon.diaz@esdeg.edu.co

Henderson elberto Rodríguez
Principios de inteligencia artificial explicable (XAI), en el diseño ético de soluciones de ciberdefensa en las Fuerzas militares colombianas, Escuela Superior de Guerra "General Rafael Reyes Prieto", Colombia. Ingeniero Electrónico, Especialista en gerencia de Mantenimiento, MBA. Alumno CIM2025, Escuela Superior de Guerra "General Rafael Reyes Prieto", Colombia. https://orcid.org/0009-0009-8838-7062.
Contacto: hendersson.rodriguez@esdeg.edu.co

Resumen

Este capítulo analiza cómo la Inteligencia Artificial Explicable (XAI) puede fortalecer la ética y la transparencia en la ciberdefensa de las Fuerzas Militares. A partir de una revisión cualitativa de literatura especializada y marcos normativos, se identificaron principios clave como la transparencia, la auditabilidad, el control humano significativo y la necesidad de explicaciones adaptadas a distintos niveles jerárquicos. Los resultados muestran que la explicabilidad no solo mejora la confianza en la tecnología, sino que también legitima la toma de decisiones militares en contextos críticos. Las conclusiones subrayan que la XAI debe ser vista como un soporte ético-técnico que equilibra rendimiento y claridad, integra al humano en el centro del proceso y requiere normas verificables para su aplicación. El estudio reconoce limitaciones por su enfoque teórico y recomienda avanzar hacia pilotos prácticos, métricas estandarizadas y una cultura institucional que garantice la adopción efectiva de estos principios.

Palabras Clave: Ciberdefensa, Fuerzas Militares FFMM; Inteligencia Artificial Explicable XAI; Explicabilidad; NATO DEEP eAcademy

Introducción

La incorporación de tecnologías de inteligencia artificial en ámbitos considerados como estratégicos, como lo es la seguridad y defensa, en la actualidad, presenta desafíos éticos y técnicos que requieren atención urgente. Uno de los problemas a los que se enfrentan las entidades del Estado e investigadores es la limitada capacidad de los sistemas de IA para proporcionar explicaciones legibles sobre sus decisiones, lo que fomenta la incertidumbre, reducción de la confianza y atenta contra la rendición de cuentas en situaciones operativas delicadas. La situación es más extrema en el ámbito militar ya que las decisiones automatizadas tienen un impacto considerable, puesto que influyen directamente en la vida de los seres humanos, en la seguridad nacional y en la correcta implementación del derecho internacional humanitario. La opacidad algorítmica, según lo expresado por O'Neil (2016) y Eubanks (2018), tiene el potencial de perpetuar violencia y conflictos, eliminar de proceso a las poblaciones más vulnerables y debilitar instituciones democráticas. Este trabajo ha

recopilado una serie de principios rectores que deberían integrarse en las operaciones de ciberdefensa ética basadas en la inteligencia artificial explicable en las fuerzas militares colombianas, teniendo en cuenta la existencia de una brecha considerable entre las suposiciones teóricas sobre la explicabilidad y las soluciones prácticas para llenar estos vacíos operacionales (Gunning & Aha, 2019; Guidotti et al., 2021).

Con base en anterior, se planteó la siguiente pregunta de investigación. ¿Cuáles deben ser los principios fundamentales de la inteligencia artificial explicable para el diseño ético de soluciones de ciberdefensa en las Fuerzas Militares colombianas? Como respuesta, esta investigación persiguió el objetivo general de determinar los principios fundamentales de la inteligencia artificial explicable (XAI) en el diseño ético de soluciones de ciberdefensa en las Fuerzas militares colombianas. Para cumplir con ese objetivo, los objetivos específicos de la investigación fueron: 1) Identificar los principios teóricos y normativos que sustentan la Inteligencia Artificial Explicable (XAI), con énfasis en su aplicación a contextos de ciberdefensa; 2) Analizar las Teorías y Normatividad sobre el uso ético de sistemas XAI en aplicaciones de ciberdefensa en las Fuerzas Militares colombianas y 3) Proponer lineamientos basados en los principios de la Inteligencia Artificial Explicable (XAI) para el diseño ético de soluciones de ciberdefensa en las Fuerzas Militares colombianas.

La metodología empleada fue un enfoque cualitativo basado en la revisión de documentos a través del análisis temático y comparativo de fuentes académicas. En primer lugar, se realizó una revisión de literatura para determinar los principios conceptuales y taxonómicos relacionados con la XAI. Segundo, se establecieron los marcos teórico y normativo nacional, buscando preceptos éticos y regulaciones técnicas para ser adaptadas a las operaciones militares colombianas y finalmente se establecieron algunos lineamientos útiles para la regulación de la XAI en el sector defensa.

En su parte teórica, esta investigación está basada en la Teoría Psicológica de la Explicabilidad, desarrollada por Yang, Folke y Shafto (2022). Esta teoría observa que las explicaciones generadas por la IA son interpretadas por los usuarios humanos con base a sus modelos mentales y expectativas cognitivas. Esta perspectiva permitió conceptualizar la explicabilidad no como una propiedad técnica, sino como un proceso de alineación entre el razonamiento del sistema y los procesos cognitivos del humano. Además, vienen a añadirse

enfoques como la lógica difusa (Alonso et al., 2021), las teorías de evaluación de calidad de la explicación (Vilone & Longo, 2021), y los marcos argumentativos aplicados a contextos jerárquicos (Demollin et al., 2020), que enriquecen este espacio con una diversidad de conceptualizaciones.

En vista de lo anterior, resulta relevante esta investigación, ante la ausencia de soluciones integrales que garanticen comprensibilidad, confiabilidad y auditabilidad de los sistemas de IA en entornos militares, pese a la existencia de propuestas teóricas sobre la materia. Además, aun cuando Colombia ha avanzado en la formulación de un marco ético para la IA, este carece de especificidad sectorial que responda a los desafíos técnicos, normativos y operacionales de las Fuerzas Militares. Por lo tanto, el propósito de este trabajo es el diseño de una arquitectura ética aplicable, contextualizada al ámbito nacional, que promueva el significativo control humano y la responsabilidad institucional dada las tecnologías emergentes para la ciberdefensa.

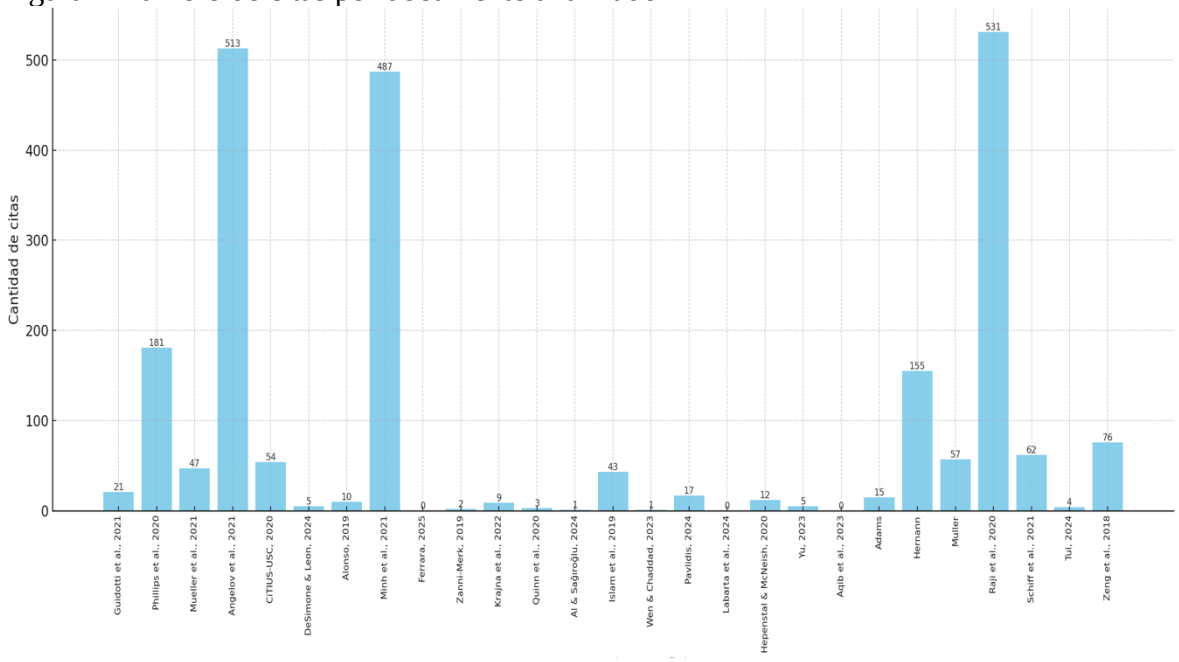
Análisis documental: Estableciendo un concepto de Inteligencia Artificial Explicable (XAI)

Para desarrollar el análisis del primer objetivo, se realizó una revisión documental de 26 fuentes académicas relacionadas con la Inteligencia Artificial Explicable (XAI). Este análisis se apoyó en los principios metodológicos propuestos por Aria y Cuccurullo (2017), en el que se priorizó el uso de métricas como el número de citas por documento para establecer el impacto relativo de cada fuente, además de considerar su pertinencia temática, año de publicación y enfoque disciplinar.

En este sentido, la selección de documentos se fundamentó en un modelo de análisis de productividad por autor y relevancia por impacto, siendo el número de citas la variable principal. Tal como explica Aria y Cuccurullo (2017), uno de los ejes centrales de este tipo de análisis es la identificación de las fuentes más influyentes mediante el recuento de citas, lo que permite establecer una base sólida y representativa del estado del arte en un campo emergente. Esta lógica fue aplicada de forma rigurosa al corpus documental sobre XAI, permitiendo distinguir las obras con mayor influencia conceptual en el área.

Además, se aplicaron estrategias de codificación temática para agrupar los documentos según categorías emergentes como ética normativa, gobernanza institucional, taxonomía explicativa y marco regulatorio. Estas técnicas, facilitaron una lectura estructurada de las contribuciones analizadas, lo que a su vez fortaleció la comprensión de los patrones conceptuales predominantes en la literatura reciente.

Figura 1 Número de citas por documento analizado



Fuente: elaboración propia

Los documentos seleccionados para el análisis conceptual muestran diversidad temática, y su distribución permite identificar cinco grandes categorías emergentes según la codificación aplicada, de acuerdo con la tabla 1:

Tabla 1 diversidad temática

Categoría temática	Documentos destacados	Rango de citas
Ética y principios normativos	Adams (15), Pavlidis (17), Zeng et al. (76),	15–76
Diseño humano-céntrico y educativo	Muller (47), Alonso (54)	47–54
Modelos técnicos y taxonomías explicativas	Angelov (513), Islam (43), Minh (487),	43–513

Categoría temática	Documentos destacados	Rango de citas
Evaluación crítica de marcos institucionales	Schiff et al. (62), Raji et al. (531)	62–531
Aplicaciones específicas y casos sectoriales	Guidotti (21), Hernann (155), Mueller (47), Phillips (181)	21–181

Fuente: elaboración propia

Ahora bien, los criterio de exclusión se fundamentaron en los documentos que presentaron baja citación (menos de 5 citas) y escasa relación con los ejes conceptuales del objetivo expresados en la tabla 2:

Tabla 2 Documentos excluidos

TITULO	CITAS
Leveraging Explainable AI in Business and Further – DeSimone & Leon, (2024)	5
Explainable AI and Mathematics – Ferrara, (2025)	0
On the Need fan Explainable Artificial Intelligence – Zanni-Merk, (2019)	2
Explainable Artificial Intelligence: An Updated Perspective – Krajna et al., (2022)	9
Recommended Methods for Using the 2020 NIST Principles for AI Explainability – Quinn et al., (2020)	3
A Review of Explainable Artificial Intelligence – Al & Sağıroğlu, (2024)	1
The Use of Explainable Artificial Intelligence in Medicine – Wen & Chaddad, (2023)	1
Study on the Helpfulness of Explainable Artificial Intelligence – Labarta et al., (2024)	0
Explainable Artificial Intelligence: What Do You Need to Know? – Hepenstal & McNeish, (2020)	12
Towards Trustworthy and Understandable AI: Explainability Strategies – Yu, (2023)	5
Advancing Trustworthy Explainable Artificial Intelligence – Aqib et al., 2023	0
Ethical Principles of Artificial Intelligence. Ethics and Information Technology. Tul, D. (2024).	4

Fuente: elaboración propia

En consecuencia, a continuación se presenta el análisis de cada uno de los documentos seleccionados:

El concepto de Inteligencia Artificial Explicable (XAI) nace como respuesta crítica al uso de modelos opacos, también conocidos como "cajas negras", en contextos donde las decisiones deben ser auditables y comprensibles. Guidotti et al. (2021) definen XAI como una disciplina centrada en la generación de explicaciones comprensibles para distintos tipos de usuarios, buscando transparencia sin comprometer el rendimiento del sistema. Esta perspectiva es clave para fundamentar la legitimidad de decisiones automatizadas en sectores críticos como justicia o defensa. La necesidad de explicar cómo se llega a una decisión técnica, con lenguaje asequible y adaptado al contexto, introduce una dimensión epistemológica y comunicativa en el desarrollo de algoritmos. Por ello, entender qué es XAI exige un enfoque interdisciplinario que incorpore no solo la computación, sino también la ética, el derecho y la psicología cognitiva.

Por su parte, Hermann (2021) ofrece un análisis aplicado de los principios de explicabilidad en el ámbito del marketing ético, mostrando cómo los valores asociados a XAI se traducen en prácticas empresariales concretas. Aunque se refiere a un entorno comercial, los principios son extrapolables a cualquier campo donde se pretenda garantizar decisiones justas y comprensibles. El autor subraya las tensiones entre rendimiento predictivo y transparencia, y plantea la necesidad de soluciones híbridas. Esta reflexión es importante para contextualizar el concepto de XAI más allá de la teoría, identificando cómo se operacionaliza en escenarios reales. Además, permite dimensionar los retos que enfrentan los diseñadores al balancear precisión, comprensión y eficiencia. Este equilibrio es crucial en sectores como defensa, donde el fallo o la incomprensión pueden acarrear consecuencias críticas.

Así mismo, Adams (2023) contribuye al marco conceptual al proponer que la explicabilidad debe ser reconocida como un principio ético autónomo dentro del conjunto de principios bioéticos tradicionales. Al defender la incorporación formal de la explicabilidad junto a principios como justicia, beneficencia o autonomía, Adams destaca que sin comprensión no puede haber consentimiento ni responsabilidad plena. Este argumento es particularmente potente en el contexto colombiano, donde decisiones automatizadas en justicia o salud requieren ser explicadas para que puedan ser legítimamente aceptadas o impugnadas. Su enfoque también posiciona a la explicabilidad como herramienta habilitadora del resto de principios, facilitando el diálogo entre la ética normativa y el diseño

técnico. Así, la definición de XAI adquiere mayor profundidad al estar anclada en discusiones filosóficas sobre la transparencia como valor moral.

Aunado a esto, el trabajo de Zeng et al. (2018) sobre la plataforma LAIP resalta que ningún conjunto de principios sobre XAI es completo por sí solo, por lo que proponen una integración de diversos marcos éticos en un sistema coherente. Esta propuesta añade una capa conceptual sobre la gobernanza ética de la explicabilidad, destacando que no solo se trata de explicar, sino de consensuar colectivamente cómo, a quién y con qué fines deben ofrecerse esas explicaciones. De este modo, la definición de XAI se amplía desde una función técnica hasta un instrumento de deliberación democrática. En este sentido, XAI no es un fin en sí mismo, sino una mediación ética entre el funcionamiento de la tecnología y los valores sociales que se desean preservar.

Continuando con la discusión sobre el papel social de la explicabilidad, el trabajo de Schiff et al. (2021) presenta una revisión comparativa entre el uso de principios éticos de inteligencia artificial en el sector público, privado y en ONGs, destacando que el entendimiento del concepto de XAI cambia según el contexto institucional. Esto implica que no existe una definición universal o fija de explicabilidad, sino una construcción contextual que depende de intereses, valores y responsabilidades particulares. Esta afirmación es clave para los sectores de defensa y justicia, donde los sistemas explicables deben articularse con normativas nacionales, cultura organizacional y preocupaciones de seguridad. De este modo, el análisis conceptual de la XAI no puede limitarse a la técnica, sino que debe considerar variables estructurales e institucionales que condicionan su implementación real.

A su vez, Raji et al. (2020) introducen el concepto de "auditoría algorítmica" como un componente complementario al de explicabilidad, subrayando que no basta con comprender un sistema, sino que también es necesario contar con herramientas para examinar sus impactos éticos y sociales. Esta perspectiva amplía la noción de XAI, vinculándola con mecanismos de control, rendición de cuentas y supervisión. En consecuencia, identificar los principios fundamentales de la IA explicable no solo requiere entender cómo se genera una explicación, sino también cómo se usa esa explicación para tomar decisiones éticas, corregir errores y generar confianza institucional. Esta visión proactiva se alinea con los retos de

ciberdefensa, donde la trazabilidad y la justificación de cada acción automatizada pueden tener implicaciones políticas, jurídicas o incluso humanitarias.

Por otra parte, el trabajo de Pavlidis (2024) ofrece una mirada desde el derecho, enfocándose en cómo el concepto de explicabilidad se integra en el marco regulatorio europeo, particularmente en la propuesta del Acta de Inteligencia Artificial de la Unión Europea. Esta perspectiva resalta que, para que un sistema sea considerado ético y legalmente aceptable, debe explicar sus procesos de decisión de forma comprensible, precisa y coherente con las expectativas normativas. Así, la explicabilidad se convierte en una obligación jurídica, no solo técnica o ética. Esta dimensión normativa es esencial al analizar el concepto de XAI en contextos institucionales, especialmente en el sector público y la administración de justicia, donde el cumplimiento legal y el debido proceso exigen claridad sobre cómo se toman las decisiones automatizadas.

Adicionalmente, el trabajo de Nguyen et al. (2023) organiza las estrategias de XAI en función del tipo de modelo, el dominio de aplicación y las necesidades del usuario, proponiendo una taxonomía funcional que facilita su clasificación y evaluación. Esta propuesta sintetiza años de investigación sobre XAI, permitiendo distinguir entre métodos intrínsecos, modelos interpretable por diseño y técnicas post hoc. Al mismo tiempo, ofrece una visión estratégica de cómo se deben adaptar los métodos explicativos a los requerimientos éticos y funcionales de cada sector. En este sentido, identificar los conceptos clave de la XAI implica también clasificar sus métodos, reconocer sus límites y analizar su aplicabilidad diferencial. Esta visión estructurada resulta esencial para el diseño de políticas y soluciones prácticas basadas en IA explicable.

Desde una perspectiva normativa, Phillips et al. (2020) proponen cuatro principios fundamentales para un sistema de IA explicable: explicación, comprensibilidad, precisión explicativa y límites del conocimiento. Estos principios son especialmente relevantes en sectores estratégicos donde los errores pueden tener consecuencias graves, pues no solo exigen que los sistemas expliquen sus decisiones, sino que lo hagan de forma verificable, inteligible y dentro de los márgenes seguros de operación. En contextos como defensa, este marco proporciona una base para auditar, confiar y supervisar el comportamiento de sistemas complejos.

Por otra parte, Angelov et al. (2021) destacan que los métodos actuales de XAI pueden clasificarse según su enfoque técnico, ya sea intrínseco o post hoc, y advierten que muchas explicaciones actuales son poco fiables o incluso engañosas. Este llamado de atención subraya la importancia de establecer marcos estandarizados y rigurosos que eviten que la "explicabilidad" se use como simple retórica tecnológica. Para entornos como la justicia, donde una mala explicación puede afectar derechos fundamentales, resulta urgente fortalecer la validez y consistencia de las técnicas utilizadas.

Desde una perspectiva educativa, Alonso et al. (2020) enfatizan que la formación en XAI debe comenzar desde etapas tempranas, integrando competencias técnicas, éticas y comunicativas. Esta aproximación se vuelve estratégica si se desea crear una cultura institucional en sectores como defensa o justicia, donde la comprensión de los sistemas automatizados no debe limitarse a los técnicos, sino extenderse a operadores, supervisores y decisores. Promover esta alfabetización explicativa puede convertirse en una medida preventiva ante el mal uso de la IA.

Por otra parte, en un esfuerzo por medir la explicabilidad, Islam et al. (2019) proponen métricas que permiten cuantificar qué tan comprensible y fiel es una explicación, con base en criterios de precisión, utilidad y tiempo de procesamiento. Estas métricas resultan útiles para organismos reguladores y operadores de sistemas críticos, ya que permiten evaluar si una explicación es efectiva más allá de su simple presencia textual o visual. Aplicar estos indicadores en defensa o justicia puede mejorar la supervisión y validación de decisiones automatizadas.

También, el trabajo de Minh et al. (2021) ofrece una revisión exhaustiva sobre el estado del arte en XAI, abordando desde fundamentos conceptuales hasta aplicaciones industriales y gubernamentales. Esta panorámica permite identificar buenas prácticas, marcos de referencia y retos comunes, lo que facilita adaptar soluciones XAI a contextos específicos. Su análisis permite también comparar la madurez tecnológica entre sectores, ubicando al ámbito judicial y militar como campos donde la explicabilidad no solo es deseable, sino necesaria para garantizar derechos, seguridad y legitimidad institucional.

Conceptos claves

Entonces, tras el anterior análisis documental se puede afirmar que la Inteligencia Artificial Explicable (XAI) constituye un eje crítico para garantizar la transparencia, la legitimidad y la ética en los sistemas automatizados de toma de decisiones. Su desarrollo responde a la necesidad de transformar modelos complejos y opacos en estructuras comprensibles, auditables y justificables, particularmente en ámbitos donde las decisiones tienen un impacto directo sobre derechos fundamentales. La XAI no es únicamente una herramienta técnica, sino un enfoque interdisciplinario que articula dimensiones normativas, comunicativas y funcionales, integrando saberes de la computación, el derecho, la ética y la psicología cognitiva. Este enfoque permite generar explicaciones adaptadas a distintos perfiles de usuarios, fortaleciendo la trazabilidad de los procesos y facilitando el control institucional. Así, la explicabilidad se convierte en un principio que no solo habilita la confianza en la tecnología, sino que también reconfigura los marcos de responsabilidad y deliberación democrática en sectores como la justicia, la salud o la defensa. En consecuencia, los conceptos claves de XAI implican reconocer su papel estructural como garante de decisiones automatizadas legítimas y coherentes con los valores sociales.

A su vez, se identificaron algunos ejes temáticos considerados como conceptos claves para la XAI de acuerdo a la frecuencia de codificación presente en la siguiente tabla

Tabla 3 Frecuencia de Codificación

Nodo / Código	Nº de referencias	Nº de fuentes
Transparencia y auditabilidad	5	3
Interdisciplinariedad	4	2
Tensión rendimiento/explicabilidad	6	3
Aplicación práctica	3	2
Explicabilidad como principio ético	4	2
Consentimiento informado	3	1
Gobernanza ética	2	1
Deliberación democrática	3	2
Contextualización institucional	3	2

Auditoría algorítmica	3	1
Rendición de cuentas	2	1
Obligación legal	3	2
Taxonomía funcional de XAI	5	2
Intrínseco vs Post hoc	4	2
Principios técnicos XAI	5	2
Fiabilidad de métodos actuales	2	1
Formación en XAI	3	1
Evaluación de explicaciones	3	2
Madurez sectorial de la XAI	2	2

Fuente: elaboración propia

A continuación se describen los conceptos claves producto de la codificación y análisis documenta:

Transparencia y auditabilidad: La XAI promueve sistemas cuyas decisiones puedan ser comprendidas y verificadas por humanos, facilitando la revisión y rastreo de procesos algorítmicos.

Interdisciplinariedad: El desarrollo de XAI requiere la integración de conocimientos provenientes de la informática, la ética, el derecho, la psicología y la sociología, entre otros campos.

Tensión rendimiento/explicabilidad: Existe un conflicto entre la precisión técnica de los modelos complejos y la necesidad de generar explicaciones simples y comprensibles para los usuarios.

Aplicación práctica: Los principios de la XAI deben ser traducidos en acciones concretas en sectores como justicia, salud, defensa o marketing ético, más allá del plano teórico.

Explicabilidad como principio ético: Se considera que explicar las decisiones de los sistemas automatizados es un deber moral, al nivel de otros principios bioéticos como la justicia o la autonomía.

Consentimiento informado: En contextos como la salud o la justicia, una explicación clara es indispensable para que los usuarios comprendan los efectos de una decisión y puedan aceptarla o rechazarla.

Gobernanza ética: La XAI debe estar guiada por principios éticos colectivos que regulen el diseño, uso y evaluación de las explicaciones ofrecidas por los sistemas inteligentes.

Deliberación democrática: La explicabilidad no debe ser solo técnica, sino que debe responder a acuerdos sociales sobre cómo y por qué se explican las decisiones automatizadas.

Contextualización institucional: La definición y aplicación de XAI varía según el marco institucional (público, privado, ONG), adaptándose a normas, culturas organizacionales y objetivos específicos.

Auditoría algorítmica: Complementa la explicabilidad al permitir revisar, controlar y corregir los efectos de los algoritmos, fortaleciendo la transparencia y confianza pública.

Rendición de cuentas: XAI fortalece la trazabilidad de las decisiones, haciendo posible identificar responsables y establecer mecanismos de supervisión y corrección de errores.

Obligación legal: En algunas jurisdicciones, como la Unión Europea, los sistemas de IA deben ofrecer explicaciones comprensibles como requisito legal de legitimidad y cumplimiento normativo.

Taxonomía funcional de XAI: Clasifica los enfoques explicativos según el tipo de modelo, aplicación y necesidades del usuario, facilitando su selección e implementación adecuada.

Intrínseco vs Post hoc: Distingue entre modelos explicables por diseño (intrínsecos) y aquellos que requieren métodos adicionales para generar explicaciones después del proceso (post hoc).

Principios técnicos XAI: Incluyen criterios como comprensibilidad, fidelidad, precisión y límites de conocimiento, que orientan el diseño y evaluación de las explicaciones generadas.

Fiabilidad de métodos actuales: Se ha señalado que muchas técnicas de XAI actuales son inestables o engañosas, lo que exige establecer estándares más sólidos y verificables.

Formación en XAI: Es clave promover la alfabetización explicativa en distintos niveles institucionales para que la comprensión de sistemas automatizados no se limite a expertos.

Evaluación de explicaciones: Incorpora métricas objetivas (precisión, utilidad, tiempo) que permiten medir la calidad y efectividad real de una explicación, más allá de su forma.

Madurez sectorial de la XAI: El grado de desarrollo de XAI varía según el sector; justicia y defensa presentan necesidades más urgentes de explicabilidad que otras industrias.

Fundamentos teóricos y desafíos normativos del uso ético de sistemas XAI en el ámbito de la ciberdefensa militar colombiana

Teorías de Inteligencia Artificial

Desde una perspectiva de análisis de teorías, se observa que los marcos teóricos que sustentan la integración de la Inteligencia Artificial Explicable (XIA) a los sistemas de ciberdefensa, son practicas teorías que fundamentan su interacción. Por ejemplo, la Teoría General de Sistemas es relevante para interpretar la ciberdefensa como un sistema complejo y adaptativo. En este sistema, cada componente, algoritmos, usuarios, reguladores, interactúa de manera dinámica y necesita coordinación sistémica. Aplicada a la XIA, facilita la comprensión estructural de los sistemas defensivos y posibilita modelar cómo las explicaciones generadas por IA impactan la toma de decisiones en contextos militares, donde la rapidez, trazabilidad e interoperabilidad son críticas (Martín, 2024). Así entendida, la ciberdefensa como un sistema integrado permite abordar la explicabilidad no como un elemento aislado, sino como un nodo central que articula el funcionamiento de varios subsistemas tecnológicos y humanos.

Lo anterior, se traduce en las Fuerzas Militares colombianas en el diseño de arquitecturas operativas donde los niveles estratégicos, tácticos y técnicos están alineados por explicaciones comprensibles que no interrumpen la cadena de mando, pero refuerzan la confianza institucional y la eficacia operacional.

Asimismo, la Teoría del empujón (Nudge Theory) puede ser considerada en este aspecto, puesto que introduce una visión más comportamental de la ciberdefensa basada en IA. Esta teoría sugiere que las explicaciones generadas por los sistemas XAI se diseñan para influir de manera positiva en las decisiones de los operadores humanos, y no para sustituirlas en la toma de decisiones (Martín, 2024). En escenarios militares, donde las recomendaciones automatizadas pueden tener consecuencias letales, la XAI no solo informa, sino que también guía conductas sutilmente hacia la reducción de riesgos cibernéticos. Asimismo, esta noción resulta especialmente útil en entornos de alta carga cognitiva y de estrés, como lo son las operaciones militares. La posibilidad de que un sistema XAI pueda operar como un "facilitador ético" y no únicamente como herramienta técnica, convierte al rol de la inteligencia artificial en algo radicalmente diferente al proceso anterior.

Por lo tanto, las capacitades de las organizaciones militares se beneficiarían no sólo en términos de eficiencia en la respuesta, sino en términos de aseguramiento a que esta respuesta estuviera habilitada por aquellos principios humanitarios y de legalidad en un guion predeterminado, sin la necesidad de intervención coercitiva.

Ahora bien, para entender qué tipo de explicaciones le es posible ofrecer a los seres humanos un sistema de inteligencia artificial y cuáles no, la teoría de la computabilidad, planteada a partir de los trabajos de Turing, Gödel y Church, (citado por Martín, 2024) se presenta como base conceptual. Desde esa perspectiva se entiende que no todas las decisiones pueden explicarse de manera computacional en términos comprensibles para los seres humanos, especialmente al enfrentar a una deep neural network (DNN. Por su sigla en inglés, red neuronal artificial) u otros modelos de caja negra. Esto plantea un desafío ético relevante para la ciberdefensa, donde sobre decisiones algorítmicas se exige rendición de cuentas, incluso en circunstancias de emergencia nacional o ciberataque crítico. Este cambio de paradigma trae a repensar, como primer punto, la ubérrima expectativa de "transparencia total" en sistemas de defensa. En tanto es una premisa que hay límites estructurales para la explicabilidad, el ejercicio en las organizaciones militares implica establecer un límite mínimo aceptable para la comprensión operativa, priorizando aquellos problemas a los que la explicación pueda dar una respuesta útil, pero no necesariamente exhaustiva: más que

buscar transpariencia absoluta, se vuelve estratégico definir el umbral de explicabilidad funcional que garantice la legítimidad, eficacia y el desempeño especulado.

Por otro lado, la teoría de la información de Shannon y Weaver (citado por Martín, 2024) también proporciona bases fundamentales para la estructuración de la comunicación entre sistemas XAI y seres humanos. Si el contexto de ciberdefensa, genera un ruido semántico claramente puede dificultar la interpretación de alertas o recomendaciones, por eso, es vital optimizar la transferencia de información explicativa. Esta teoría da pautas para analizar cómo la entropía o la redundancia informativa modulan la calidad de las explicaciones, participando así en el diseño de interfaces y estructuras comunicativas en sistemas de seguridad y defensa nacional en situaciones particulares en las que estas explicaciones deben adentrarse en entornos burocratizado. Este enfoque conlleva un avance hacia el diseño centrado en el usuario militar, donde la explicabilidad no es tanto una propiedad del sistema como una experiencia ajustada a un contexto.

Por ende en Colombia, donde deben coexistir entornos de despliegue flexibles con múltiples niveles de conocimiento técnico, estos principios permitirían el diseño de plataformas explicativas personalizadas por rango, función o rol, asegurando así que la información no solo llegue sino que sea entendida y en efecto, útil para la acción. Por tanto, se debe enfatizar que todas las teorias convergen en la necesidad de considerar la IA explicable no solo como un componente técnico, sino como una arquitectura conceptual que articula tecnología, comunicación, comportamiento y sistemas complejos. En el caso específico de las Fuerzas Militares, implica considerar la pertinencia de adaptar dichos marcos teóricos a un entorno institucional específico, donde la doctrina, el secreto y la autonomía para dictar juicios requieran hacer de las explicaciones una experiencia ajustada a la realidad operacional, legal y cultural. Siendo así, proponer una ruta metodológica para la integración de XAI en defensa que parta no solo del análisis técnico, sino de un enfoque integral que valore la dimensión organizacional, ética y pedagógica de la explicabilidad, se convierte en un reto para los programadores del sector defensa, ya que no se trata de aplicar ciegamente marcos universales, sino de reinterpretarlos a la luz de las amenazas, estructuras y valores propios de las Fuerzas Militares. En consecuencia, la implementación efectiva de

la XAI exige tanto innovación técnica como reflexión institucional, construida sobre bases teóricas sólidas y contextualizadas.

Teoría Psicológica de la Explicabilidad Como Ente Integrador

La intersección entre la Teoría Psicológica de la Explicabilidad de Yang, Folke y Shafto (2022) y los marcos sistemáticos revisados en los párrafos anteriores revela una teorización complementaria que refuerza la centralidad del usuario en la arquitectura explicativa de los sistemas XAI aplicables a la ciberdefensa. En tanto la teoría psicológica dice relación con la necesidad de que las explicaciones se alineen con los modelos mentales y expectativas pasadas de los usuarios, teorías como el Empujón y la Teoría de la Información de Shannon y Weaver resaltan la forma en que esa explicación misma debe estructurarse y presentarse para obtener un efecto conductual claro en la toma de decisiones operacionales. En ese sentido, se establece una continuidad entre la comprensión individual de la explicación y su función instrumental dentro de sistemas de defensa complejos, en una línea argumentativa continua con la Teoría General de Sistemas.

Esta relación resulta especialmente valiosa en entornos militares jerárquicos, donde la comprensión de una explicación no solo debe ser cognitivamente exitosa, sino también efectiva comunicativamente y ética y funcionalmente compatible con los procedimientos ya establecidos. Por su parte, las contribuciones de Demollin et al. (2020) y de Thompson (2018) también extienden esta comprensión, al entregar marcos de argumentación y unificación que relacionan la calidad de la explicación con su adaptabilidad contextual institutional, fortaleciendo la aseveración de que el "explicar" en IA no es tan solo una función técnica, sino, más que nada, socialmente situada. Fruto de esta articulación teórica, la Teoría Psicológica opera como eje de integración que otorga sentido a los demás marcos conceptuales al situar el acto explicativo como acto comunicativo y ético receptivo. Dada su operación en contextos de máxima incertidumbre, la ciberdefensa no puede permitirse el lujo de depender de sistemas cuya explicabilidad sea medida por su capacidad de expresar lógicas internas, sino que de hecho debe preocupar su capacidad de generar interpretaciones, acciones y, en última estancia, confianza humana. Descriptores como "fiabilidad", "claridad"

y “control humano significativo” extraídos de Wang et al. (2019), Moral et al (2021), y Vilone & Longo (2021), encuentran un aterrizaje validante en la psicología de la explicación, que contextualiza al usuario no como receptor sino como actor interpretativo. La conjunción entre estas teorías implica que las soluciones XAI, deben ser tanto autónomas y robustas tecnológicamente como epistémicamente inclusivas en términos cognitivos, funcionales y culturales, promoviendo un modelo explicativo anclado tanto de la técnica como de la comprensión humana.

Normatividad nacional sobre inteligencia artificial

El entorno jurídico que acompaña a la inteligencia artificial en Colombia se encuentra en una etapa de consolidación. Aunque se han establecido ciertos principios rectores, como la transparencia, la responsabilidad y la protección de los derechos humanos, no existe un marco legal completo que regule adecuadamente el uso ético de sistemas de inteligencia artificial explicables en sectores altamente sensibles, como la ciberdefensa. Dada la ausencia de legislación, una de las alternativas es adoptar el marco ético de inteligencia artificial de Colombia (MinTIC, 2022). Si bien no tiene carácter vinculante, contiene cinco principios rectores que deben guiar el desarrollo, la aplicación y la supervisión de cualquier sistema de IA, y uno de ellos se centra en la explicabilidad. La explicabilidad se considera necesaria porque garantiza la confianza en el estado de derecho y el derecho a la autonomía de los usuarios finales.

De este modo, el proyecto de Ley 43 de 2025 (Congreso de la República de Colombia, 2025) es un paso en la formalización de una regulación con pretensión jurídica, toda vez que define obligaciones específicas para los actores que desarrollen o implementen sistemas de IA. Sin embargo, este instrumento aún no se ha convertido en ley, y las tensiones que genera evidencian que las legislaciones tradicionales no son fácilmente adaptables a estas tecnologías. La discusión parlamentaria no se centra simplemente en los alcances técnicos de la regulación, sino, en paralelo, en la preservación de valores democráticos, soberanía de los datos y responsabilidad tripartita entre desarrolladores, operadores e instituciones del Estado.

Es, por ende, una zona de transición, más marcada por recomendaciones normativas de alto contenido ético-político que por normativas de cumplimiento. Esto genera dudas

sobre los mecanismos prácticos de control, trazabilidad y auditoría sobre bases algorítmicas permitidos ante un despliegue operativo de sistemas XAI en el ámbito de la defensa. Así, el reto de la normativa XAI no es solo proponer principios de alto nivel, sino operacionalizarlos en contextos institucionales como las Fuerzas Militares, donde las decisiones automatizadas pueden tener consecuencias políticas, legales o humanitarias. Por ende, la aplicabilidad práctica de la explicabilidad, la transparencia y la rendición de cuentas no es solo un problema de discusión sino una necesidad en la regulación nacional en cuanto a su uso.

Como lo han advertido varios documentos oficiales y académicos (MinTIC, 2024; Pezzini & Pons, 2024), una IA que decide sin dejar rastro comprensible de su lógica vulnera no solo el principio de debido proceso, sino también los fundamentos de la ética pública. En este sentido, se vuelve imprescindible contar con marcos normativos, que consideren tanto el rápido avance tecnológico como la necesidad de preservar la responsabilidad humana en escenarios automatizados.

Entonces, cuando se aplica al campo de la ciberdefensa, este marco inacabado presenta peligros y oportunidades. En un lado, la falta de una normativa específica podría dar lugar a desarrollo descontrolado, debilitando las garantías institucionales y arrojando duda sobre la ciudadanía. En el otro, se abre la posibilidad de una regulación realizada con una perspectiva de intervención estratégica, que entrecruza los diferentes desarrollos que componen las XAI con los principios constitucionales y los retos propios de la realidad militar colombiana. De este modo, urge avanzar hacia una normativa no solo reguladora del uso de la IA, si no que, además, enuncie específicamente la necesidad de explicabilidad como condición para el uso legítimo de estas tecnologías en el espacio de la defensa nacional.

Lineamientos Éticos para el Diseño de Soluciones de Ciberdefensa basadas en Inteligencia Artificial Explicable (XAI) en las Fuerzas Militares Colombianas

Como punto de partida, resulta imprescindible que los sistemas de inteligencia artificial desarrollados para ciberdefensa incorporen mecanismos de transparencia operacional, entendida como la capacidad del sistema para exponer de forma clara y accesible los procesos lógicos que sustentan sus decisiones, este principio de transparencia, no debe limitarse a la documentación técnica, sino traducirse en explicaciones comprensibles para operadores militares, supervisores e instancias de control, asegurando que las decisiones automatizadas puedan ser auditadas en términos éticos, técnicos y legales.

En estrecha relación con la transparencia, se requiere que las soluciones XAI integren estructuras de auditabilidad interna, que permitan reconstruir cada proceso decisional ante eventuales requerimientos institucionales o jurídicos, por eso, la auditabilidad no solo refuerza la trazabilidad del sistema, sino que respalda la responsabilidad institucional ante actores nacionales e internacionales, consolidando una cultura de rendición de cuentas alineada con los marcos normativos vigentes.

Además, debe garantizarse que los niveles de explicabilidad se adapten al perfil cognitivo y funcional de los usuarios militares, mediante el uso de modelos de explicabilidad graduada o contextualizada. Lo anterior implica que las explicaciones no serán homogéneas para todos los actores, sino que se diseñarán conforme a las necesidades de comprensión de quienes operan, supervisan o evalúan el sistema, respetando sus competencias, grados de responsabilidad y escenarios de acción.

En ese sentido, la alfabetización institucional en inteligencia artificial explicable también es un tema que cobra relevancia, pues la formación continua en ética algorítmica, lógica de decisiones automatizadas y principios de XAI debe ser integrada como parte de la doctrina militar tecnológica, a fin de empoderar al personal en el uso crítico de estas herramientas y evitar su adopción sin criterio o como una herramienta que genere dependencia.

De forma complementaria, las soluciones diseñadas deben asegurar en todo momento el control humano significativo, garantizando que los operadores conserven autoridad decisoria sobre los sistemas, en consecuencia, este principio responde a las exigencias éticas del derecho internacional humanitario y al mandato constitucional de subordinación tecnológica al juicio humano en contextos de seguridad y defensa, donde los errores pueden comprometer vidas humanas o la soberanía nacional.

Igualmente, todo diseño de sistemas XAI deberá estar en consonancia con el marco normativo colombiano sobre inteligencia artificial, así como con estándares internacionales de ética tecnológica, ya que la observancia de principios como la legalidad, la proporcionalidad, la no discriminación y el respeto por los derechos fundamentales es una exigencia que trasciende lo técnico, configurándose como fundamento de legitimidad institucional ante la ciudadanía y los organismos de control.

Asimismo, el desarrollo de soluciones XAI en contextos militares debe estar precedido por procesos rigurosos de validación, tanto en entornos simulados como operacionales, que permitan evaluar la eficacia funcional de los sistemas, así como la claridad, precisión y utilidad de las explicaciones que generan, por ende, es necesario poner a prueba los sistemas, siendo susceptibles a verificación y deberán formar parte de las etapas de diseño, despliegue y mejora continua.

En paralelo, la gestión ética de los datos adquiere una relevancia fundamental, dado que los sistemas de ciberdefensa operan con información sensible de alta clasificación, por ello, se deben establecer protocolos sólidos para el tratamiento, protección y trazabilidad de los datos, minimizando riesgos de sesgos, filtraciones o usos indebidos, y asegurando el cumplimiento del principio de minimización y proporcionalidad en el acceso y procesamiento de la información.

Además, las explicaciones ofrecidas por los sistemas XAI deben considerar el contexto operacional en el que se implementen, reconociendo las particularidades sociopolíticas, jurídicas y estratégicas del entorno militar colombiano. Esto implica que la explicabilidad no puede concebirse como un módulo técnico aislado, sino como una dimensión contextual del diseño, que dialogue con las condiciones materiales, culturales y normativas del campo de acción.

Por otra parte, los principios de deliberación democrática deben ser considerados en el diseño de estas soluciones, promoviendo espacios de participación entre diseñadores, usuarios militares, entes reguladores y actores académicos o sociales, pues estas instancias permitirán consensuar criterios éticos sobre el alcance, los límites y las finalidades del uso de la IA explicable en defensa, previniendo su instrumentalización indebida o tecnocrática.

También debe asegurarse que las soluciones XAI fomenten la corresponsabilidad institucional, promoviendo una distribución clara de funciones entre desarrolladores, operadores y autoridades de supervisión ya que se debe ser consecuente en que las decisiones algorítmicas debe servir no solo para comprender lo ocurrido, sino para asignar responsabilidades en casos de fallos o afectaciones, y para generar mecanismos correctivos oportunos.

A su vez, se recomienda incorporar indicadores cuantitativos y cualitativos de evaluación de explicaciones, tales como claridad, fidelidad, utilidad y tiempo de comprensión, que permitan medir la efectividad real de los sistemas XAI en su interacción con usuarios humanos, para ello, es importante realizar ejercicios de análisis a través de metodologías similares al proceso militar para lo toma de decisiones, en donde se realizan supuestos y se vean indicadores de evaluación, cuyas métricas no deben ser solo parte del diseño técnico, sino criterios para la toma de decisiones estratégicas sobre la continuidad, ajuste o retiro de soluciones implementadas en inteligencia artificial.

Conclusiones

1. La XAI como esqueleto ético-técnico de la ciberdefensa

El estudio confirma que sin transparencia y auditabilidad no puede existir confianza en los sistemas de inteligencia artificial aplicados a ciberdefensa, cumpliendo así con los objetivos planteados y validando la hipótesis de que la XAI sostiene el diseño ético en las FF. MM. Su significado radica en que la explicabilidad es requisito de legitimidad y no un accesorio técnico. La principal implicación es que las políticas y doctrinas militares deben exigir trazabilidad y controles de auditoría, aunque el estudio se limita a un enfoque teórico. Futuras investigaciones deberían implementar pilotos controlados para medir el impacto real de la XAI en operaciones de seguridad.

2.La persona al centro de la explicación

El documento evidencia que una explicación solo es valiosa cuando puede ser comprendida y utilizada por los diferentes roles humanos que toman decisiones, respondiendo al objetivo de articular teorías explicativas con lineamientos aplicables. Esto significa que en ambientes de alta presión la claridad cognitiva puede salvar vidas y reducir errores. La implicación directa es diseñar explicaciones diferenciadas por nivel jerárquico, aunque aún falta evidencia empírica sobre usabilidad. Se sugiere realizar estudios experimentales que midan tiempos de comprensión, confianza y carga cognitiva en mandos y operadores.

3. La normatividad como habilitadora

El análisis de marcos normativos muestra que Colombia cuenta con principios de ética en IA, pero aún no existen criterios operativos verificables en defensa, respondiendo al objetivo de contextualización legal. La interpretación es clara: se requiere pasar del “qué” al “cómo” mediante estándares, métricas y auditorías aplicables. Esto implica incluir cláusulas de explicabilidad en procesos de adquisición tecnológica y en protocolos operativos. La

limitación está en la ausencia de un modelo robusto de gobernanza con roles definidos, y las futuras investigaciones deberían diseñar y probar esquemas nacionales de conformidad en XAI para defensa.

4. El dilema rendimiento–explicabilidad

El capítulo revela la tensión entre modelos de alto desempeño y la necesidad de ofrecer explicaciones claras, lo que responde al objetivo de definir principios técnicos y lineamientos prácticos. La interpretación es que en defensa no basta con precisión matemática: se requieren modelos que también permitan justificar decisiones. La implicación es adoptar métricas duales que midan simultáneamente desempeño y calidad explicativa, aunque la investigación no fija umbrales concretos de aceptabilidad. El paso siguiente sería realizar estudios comparativos entre modelos intrínsecos y post hoc en escenarios reales de detección de ciberamenazas.

5. Control humano y corresponsabilidad institucional

La investigación sostiene que la autoridad decisoria debe permanecer en manos humanas y que las funciones de desarrolladores, operadores y supervisores deben estar claramente delimitadas, cumpliendo con el objetivo de proponer lineamientos éticos. El significado de este hallazgo es que la explicabilidad refuerza la legitimidad del mando militar y permite compatibilizar decisiones algorítmicas con el derecho internacional humanitario. La implicación es establecer protocolos claros de quién valida y documenta las recomendaciones de la IA. Sin embargo, faltan estudios de casos prácticos que evidencien cómo estas explicaciones cambian decisiones. Futuras investigaciones deberían construir un repositorio de lecciones aprendidas a partir de ejercicios y operaciones reales.

6. De la teoría a la práctica: cultura y piloto

El documento ofrece un andamiaje conceptual robusto y un conjunto de lineamientos aplicables, pero carece de pruebas empíricas, cerrando así el ciclo del tercer objetivo. El significado de esto es que la cultura organizacional en torno a la XAI es tan determinante como la tecnología misma. Las implicaciones incluyen la necesidad de formación especializada, programas de alfabetización en XAI y políticas de gobernanza de datos dentro de las FF. MM. La principal limitación es no contar con una hoja de ruta temporal definida. Como proyección, se recomienda construir un plan de implementación a 24 meses con pilotos trimestrales, métricas estandarizadas y auditorías independientes.

Referencias

Adams, J. (2023). Defending explicability as a principle for the ethics of artificial intelligence in medicine. Medicine, Health Care, and Philosophy, 26, 615 - 623. https://doi.org/10.1007/s11019-023-10175-7.

Al, S., & Sağıroğlu, Ş. (2024.). A Review of Explainable Artificial Intelligence. 2024 9th International Conference on Computer Science and Engineering (UBMK), 310-315. https://doi.org/10.1109/UBMK63289.2024.10773588.

Alonso, J. (2020). Teaching Explainable Artificial Intelligence to High School Students. Int. J. Comput. Intell. Syst., 13, 974-987. https://doi.org/10.2991/ijcis.d.200715.003.

Angelov, P., Soares, E., Jiang, R., Arnold, N., & Atkinson, P. (2021). Explainable artificial intelligence: an analytical review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11. https://doi.org/10.1002/widm.1424.

Congreso de Colombia. (2025). Proyecto de Ley No. 43 de 2025 - Por medio del cual se regula la inteligencia artificial en Colombia.

Demollin, M., Shaheen, Q., Budzynska, K., & Sierra, C. (2020). Argumentation Theoretical Frameworks for Explainable Artificial Intelligence., 44-49. https://aclanthology.org/2020.nl4xai-1.10.pdf

DeSimone, H., & Leon, M. (2024). Leveraging Explainable AI in Business and Further. 2024 IEEE Opportunity Research Scholars Symposium (ORSS), 1-6. https://doi.org/10.1109/ORSS62274.2024.10697961.

Guidotti, R., Monreale, A., Pedreschi, D., & Giannotti, F. (2021). Principles of Explainable Artificial Intelligence. Explainable AI Within the Digital Transformation and Cyber Physical Systems. https://doi.org/10.1007/978-3-030-76409-8_2.

Ferrara, M. (2025). Explainable artificial intelligence and mathematics: What lies behind? Let us focus on this new research field. European Mathematical Society Magazine. https://doi.org/10.4171/mag/235.

Hermann, E. (2021). Leveraging Artificial Intelligence in Marketing for Social Good—An Ethical Perspective. Journal of Business Ethics, 179, 43 - 61. https://doi.org/10.1007/s10551-021-04843-y.

Hepenstal, S., & McNeish, D. (2020). Explainable Artificial Intelligence: What Do You Need to Know?. , 266-275. https://doi.org/10.1007/978-3-030-50353-6_20.

Islam, S., Eberle, W., & Ghafoor, S. (2019). Towards Quantification of Explainability in Explainable Artificial Intelligence Methods. , 75-81.

Krajna, A., Kovac, M., Brčič, M., & Šarčević, A. (2022). Explainable Artificial Intelligence: An Updated Perspective. 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), 859-864. https://doi.org/10.23919/mipro55190.2022.9803681.

Labarta, T., Kulicheva, E., Froelian, R., Geissler, C., Melman, X., & Von Klitzing, J. (2024.). Study on the Helpfulness of Explainable Artificial Intelligence. ArXiv, abs/2410.11896. https://doi.org/10.1007/978-3-031-63803-9_16.

Martín, M,. (2024) Inteligencia Artificial: Un estudio de su impacto en Ciberseguridad. Universidad Abierta de Cataluña

Ministerio de Tecnologías de la Información y las Comunicaciones [MinTIC]. (2022). Marco ético para la inteligencia artificial en Colombia.

Minh, D., Wang, H., Li, Y., & Nguyen, T. (2021). Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review, 55, 3503 - 3568. https://doi.org/10.1007/s10462-021-10088-y.

Moral, J., Castiello, C., Magdalena, L., & Mencar, C. (2021). Toward Explainable Artificial Intelligence Through Fuzzy Systems. Explainable Fuzzy Systems. https://doi.org/10.1007/978-3-030-71098-9_1.

Mueller, S., Veinott, E., Hoffman, R., Klein, G., Alam, L., Mamun, T., & Clancey, W. (2021). Principles of Explanation in Human-AI Systems. ArXiv, abs/2102.04972.

Observatorio de IA y Ética Digital MinTIC. (2024). ¿En qué va la Inteligencia Artificial en Colombia?.

Pavlidis, G. (2024). Unlocking the black box: analysing the EU artificial intelligence act's framework for explainability in AI. Law, Innovation and Technology, 16, 293 - 308. https://doi.org/10.1080/17579961.2024.2313795.

Pezzini, M. C., & Pons, C. (2024). Inteligencia Artificial Explicable: Análisis de Metodologías y Aplicaciones. Universidad Nacional de La Plata.

Phillips, P., Hahn, C., Fontana, P., Broniatowski, D., & Przybocki, M. (2020). Four Principles of Explainable Artificial Intelligence. https://doi.org/10.6028/nist.ir.8312-draft.

Quinn, M., Piper, B., Bliss, J., & Keever, D. (2020). Recommended Methods for Using the 2020 NIST Principles for AI Explainability. 2020 IEEE International Conference on Big Data (Big Data), 2034-2037. https://doi.org/10.1109/BigData50022.2020.9377760.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). 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* (pp. 33-44). https://dl.acm.org/doi/pdf/10.1145/3351095.3372873

Sadiq, Z., & Aqib, M. (2023). Advancing Trustworthy Explainable Artificial Intelligence: Principles, Goals, and Strategies. OALib. https://doi.org/10.4236/oalib.1110870.

Schiff, D., Borenstein, J., Biddle, J., & Laas, K. (2021). AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection. IEEE Transactions on Technology and Society, 2, 31-42. https://doi.org/10.1109/TTS.2021.3052127.

Thompson, J. (2018). Towards a common theory of explanation for artificial and biological intelligence. . https://doi.org/10.32470/CCN.2018.1259-0.

Tul, D. (2024.). Ethical principles of artificial intelligence. XXI međunarodni naučni skup Pravnički dani - Prof. dr Slavko Carić, na temu: Odgovori pravne nauke na izazove savremenog društva - zbornik radova. https://doi.org/10.5937/pdsc24759t.

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf. Fusion, 76, 89-106. https://doi.org/10.1016/J.INFFUS.2021.05.009.

Wang, D., Yang, Q., Abdul, A., & Lim, B. (2019). Designing Theory-Driven User-Centric Explainable AI. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3290605.3300831.

Wen, B., & Chaddad, A. (2023). The Use of Explainable Artificial Intelligence in Medicine. 2023 IEEE International Conference on E-health Networking, Application & Services (Healthcom), 251-252. https://doi.org/10.1109/Healthcom56612.2023.10472365.

Yang, S., Folke, T., & Shafto, P. (2022). A psychological theory of explainability. , 25007-25021. https://doi.org/10.48550/arXiv.2205.08452.

Yu, S. (2023.). Towards Trustworthy and Understandable AI: Unraveling Explainability Strategies on Simplifying Algorithms, Appropriate Information Disclosure, and High-level Collaboration. Proceedings of the 26th International Academic Mindtrek Conference. https://doi.org/10.1145/3616961.3616965.

Zanni-Merk, C. (2019). On the Need of an Explainable Artificial Intelligence. , 3. https://doi.org/10.1007/978-3-030-30440-9_1.

Zeng, Y., Lu, E., & Huangfu, C. (2018). Linking Artificial Intelligence Principles. ArXiv, abs/1812.04814.