FACULTAD DE DERECHO

INTELIGENCIA ARTIFICIAL: UNA CUESTIÓN ÉTICA

Autora: María Dolores Fernández-Martos García-Herrera
5º E-3 B

Filosofía del Derecho

Tutor: Alberto de Unzurrunzaga Rubio

Madrid, Marzo 2025

RESUMEN

Este Trabajo de Fin de Grado explora los recientes y grandes avances en materia de inteligencia artificial (IA) y la apremiante necesidad de integrar principios éticos en su regulación con el propósito de asegurar su uso responsable. Es esencial garantizar la intervención y supervisión humana en los sistemas de IA, especialmente en áreas que impactan en los derechos fundamentales. La legislación debe equilibrar la innovación tecnológica con la protección de la autonomía y los derechos de los individuos, priorizando la transparencia y la explicabilidad en los procesos de toma de decisiones automatizadas. En el caso de España, la alineación con las directrices europeas y la integración de principios éticos en la legislación muestran el compromiso de lograr una regulación que promueva la competitividad tecnológica, pero sin sacrificar los principios fundamentales. La filosofía del derecho debe guiar el desarrollo de normativas que aseguren que la IA esté al servicio de la humanidad, y no al contrario, respetando principios tan relevantes como los de equidad, privacidad y seguridad.

Palabras clave: Inteligencia Artificial, Ética, Derechos Fundamentales, España, Unión Europea, Transparencia, Sesgo y Explicabilidad.

ABSTRACT

This paper explores recent major advances in artificial intelligence (AI) and the pressing need to integrate ethical principles into its regulation in order to ensure its responsible use. It is essential to ensure human intervention and oversight in AI systems, especially in areas that impact fundamental rights. Legislation must balance technological innovation with the protection of the autonomy and rights of individuals, prioritizing transparency and explainability in automated decision-making processes. In the case of Spain, alignment with European guidelines and the integration of ethical principles into legislation show a commitment to achieving regulation that promotes technological competitiveness, but without sacrificing fundamental principles. The philosophy of law should guide the development of regulations that ensure that AI is at the service of humanity, and not the other way around, respecting such relevant principles as fairness, privacy and security.

Key words: Artificial Intelligence, Ethics, Fundamental Rights, Spain, European Union, Transparency, Bias and Explainability.

ÍNDICE

1. INTRODUCCIÓN

1.1. Planteamiento del problema…… 6

1.2. Objetivos del estudio…… 7

1.3. Metodología……... 8

1.4. Justificación de la investigación…….. 8

2. FUNDAMENTOS DE LA INTELIGENCIA ARTIFICIAL……. 9

2.1 Definición y evolución de la IA……. 9

2.2 Aplicaciones actuales de la IA…….. 12

2.3 Desafíos y oportunidades…….. 13

3. ÉTICA EN LA INTELIGENCIA ARTIFICIAL……15

3.1 Concepto de ética en la tecnología……16

3.2 Principios éticos aplicables a la IA……... 17

3.3 Principios éticos aplicables a la IA……... 18

3.3.1 Sesgo algorítmico y discriminación……18

3.3.2 Falta de transparencia y explicabilidad……... 22

3.3.3 Privacidad y vigilancia masiva……..... 25

3.3.4 Desplazamiento laboral y automatización……... 26

3.3.5 *Uso de IA en la guerra y sistemas autónomos letales*……..27

4. REGULACIÓN JURÍDICA DE LA IA EN LA UNIÓN EUROPEA…….. 31

4.1 Análisis del Reglamento de Inteligencia Artificial……. 31

4.2 Evaluación ética del marco normativo europeo……. 39

4.3 Impacto en los Estados miembros…… 40

5. ESTRATEGIA ESPAÑOLA EN MATERIA DE IA……. 42

5.1 Objetivos y pilares de la Estrategia Nacional de IA……...42

5.2 Anteproyecto de Ley para el buen uso y la gobernanza de la Inteligencia Artificial e integración de principios éticos en la estrategia española..…….. 44

5.3 Comparación con otras estrategias nacionales……... 47

6. PERPECTIVAS INTERNACIONALES SOBRE LA ÉTICA EN LA IA... 49

6.1 Enfoque de Estados Unidos en la ética de la IA……. 49

6.2 Enfoque de China en la ética de la IA……50

7. PROPUESTA DE PRINCIPIOS ÉTICOS PARA LA REGULACIÓN JURÍDICA DE LA IA EN ESPAÑA……. 52

7.1 Identificación de principios clave……. 52

7.2 Recomendaciones para la incorporación en la legislación española…. 54

8. CONCLUSIONES……... 57

9. BIBLIOGRAFÍA……. 59

1. INTRODUCCIÓN

La Inteligencia Artificial (en adelante, IA) ha emergido como una de las tecnologías más influyentes y disruptivas del siglo XXI, transformando profundamente diversos sectores, desde la industria y el comercio hasta la sanidad, la educación y la administración pública. Su capacidad para procesar enormes volúmenes de datos y aprender patrones a través de algoritmos ha generado avances significativos en eficiencia, precisión y personalización de procesos y servicios (Brynjolfsson & McAfree, 2014). Sin embargo, el crecimiento acelerado de la IA también ha generado importantes dilemas éticos y jurídicos, relacionados con cuestiones de privacidad, discriminación, transparencia y control de la tecnología (O´Neil, 2016).

Como ha señalado la Unión Europea (Comisión Europea, 2021), la IA representa un hito en la evolución tecnológica, con el potencial de mejorar la calidad de vida y optimizar procesos en distintos ámbitos. No obstante, los organismos internacionales y los expertos coinciden en que su desarrollo y uso deben enmarcarse dentro de principios éticos bien definidos para evitar riesgos y perjuicios sociales y de otro tipo (Bryson, 2021). La IA no solo debe garantizar el respeto a los derechos fundamentales de las personas, sino que también debe ser comprensible, confiable y equitativa.

En este contexto, la ética en la IA ha cobrado un protagonismo especial en los debates académicos y normativos. En los últimos años, la comunidad internacional ha trabajado en la elaboración de principios y directrices para garantizar que la IA sea una herramienta beneficiosa para la sociedad y no una amenaza para la justicia y la equidad (Floridi et al., 2018). La Comisión Europea, por ejemplo, ha desarrollado un marco regulatorio para la IA con el fin de garantizar su uso responsable, equilibrando el impulso de la innovación con la protección de los derechos ciudadanos (Comisión Europea, 2021). A su vez, diversos países han adoptado estrategias nacionales para el desarrollo de la IA, cada uno con diferentes enfoques en materia de ética y regulación (Ministerios de Asuntos Económicos y Transformación Digital, 2021).

Veremos en uno de los próximos apartados de este trabajo que España también está dando los primeros pasos en el desarrollo normativo de esta materia, con la reciente aprobación en Consejo de Ministros del Anteproyecto de Ley para el Buen Uso y la Gobernanza de la Inteligencia Artificial, impulsado por el Ministerio para la Transformación Digital y de la Función Pública.

1.1. Planteamiento del problema

Uno de los principales desafíos que plantea la IA es su capacidad para tomar decisiones de manera autónoma, lo que genera incertidumbre sobre la responsabilidad jurídica en caso de errores o daños. ¿Quién es responsable si un algoritmo de IA discrimina a un grupo de personas en un proceso de selección laboral? ¿Quién debe responder si un vehículo autónomo causa un accidente? ¿Cómo garantizar que los sistemas de IA operen de forma transparente y explicable para los usuarios? (Gunning & Aha, 2019).

Otro aspecto crucial es la posible reproducción de sesgos en los algoritmos de IA. Dado que los modelos de aprendizaje automático se entrenan con datos históricos, pueden perpetuar y amplificar desigualdades preexistentes. Se han documentado casos en los que sistemas de IA han discriminado a mujeres, minorías étnicas y otros grupos vulnerables debido a la forma en que fueron diseñados (Buolamwini & Gebru, 2018). Por ello, es esencial evaluar cómo se puede mitigar el impacto de estos sesgos y qué mecanismos de regulación y supervisión deben implementarse para evitar injusticias y otros problemas.

Además, la IA plantea dilemas éticos en torno a la privacidad y la vigilancia masiva. En un mundo donde los datos son considerados el nuevo petróleo, muchas empresas y gobiernos utilizan IA para analizar información personal, lo que ha generado preocupaciones sobre la protección de la privacidad. Tecnologías como el reconocimiento facial y la monitorización de redes sociales han sido objeto de controversias debido a su uso potencial para el control social y la restricción de libertades individuales, especialmente en países con regímenes autoritarios (Feldstein, 2019).

Estos desafíos requieren una respuesta normativa y ética adecuada, que garantice que la IA se desarrolle dentro de límites bien definidos y que sus beneficios sean accesibles para toda la sociedad sin comprometer principios fundamentales como la dignidad, la equidad y la autonomía individual (Russell, 2019).

1.2. Objetivos del estudio

El presente trabajo tiene como objetivo analizar la ética de la IA desde la perspectiva de la Filosofía del Derecho, con el fin de delimitar los principios y valores éticos que deben trasladarse a la regulación jurídica de esta tecnología en España y otros países. Para ello, se abordarán los siguientes objetivos específicos:

a. Definir y contextualizar la ética en la IA, explorando sus principales implicaciones filosóficas, sociales y jurídicas.
b. Identificar los principios éticos clave que deberían guiar el desarrollo y uso de la IA, tales como la transparencia, la responsabilidad, la equidad y el respeto a la privacidad.
c. Analizar el marco normativo europeo sobre IA, con especial énfasis en el Reglamento Europeo de Inteligencia Artificial, recientemente aprobado por la Unión Europea.
d. Evaluar la estrategia española en materia de IA, analizando su compatibilidad con los principios éticos identificados y su eficacia en la promoción de una IA justa y equitativa.
e. Comparar el enfoque ético de la IA en Europa con los modelos adoptados por Estados Unidos y China, con el propósito de entender diferentes perspectivas y valorar la viabilidad del modelo europeo como referencia internacional.

A través de este análisis, se pretende aportar una visión integral de la relación entre la ética y la regulación de la IA, proporcionando propuestas para una legislación que garantice el desarrollo de una IA responsable y alineada con los valores democráticos.

1.3. Metodología

Para alcanzar los objetivos propuestos, el estudio se basará en un análisis documental y comparativo. Se examinarán informes y documentos relevantes sobre IA y ética, publicados por organismos como la Unión Europea, Naciones Unidas, UNESCO y la OCDE. Asimismo, se estudiará la legislación y estrategias nacionales en materia de IA en diferentes países.

El presente trabajo también incluirá una comparación entre los marcos normativos de Europa, Estados Unidos y China, evaluando sus diferencias en términos de principios éticos y regulación jurídica. Para ello, se analizarán documentos como el Reglamento Europeo de IA (Comisión Europea, 2021), la Estrategia Nacional de IA en España (Ministerio de Asuntos Económicos y Transformación Digital), la Estrategia Nacional de IA de EE.UU. y la política de IA en China (Mozur, 2019).

Además, se revisará la literatura académica existente en torno a la ética de la IA, en filosofía del derecho, derecho tecnológico y regulación de la IA.

Finalmente, se realizará una reflexión crítica sobre las implicaciones éticas de la IA en el contexto jurídico español, con el objetivo de proponer recomendaciones para una regulación más efectiva y justa.

1.4. Justificación de la investigación

El objetivo de este trabajo radica en la necesidad de establecer un marco ético y jurídico sólido para la IA, que permita maximizar sus beneficios al tiempo que se minimizan sus riesgos. Dado que la IA está impactando en todos los ámbitos de la sociedad, resulta esencial que su desarrollo y aplicación estén alineados con valores fundamentales como la dignidad humana, la equidad y la transparencia (O'Neil, 2016).

En España, la Estrategia Nacional de IA busca promover un modelo de IA responsable, sostenible y centrado en las personas. Sin embargo, aún existen desafíos en la implementación de principios éticos en la legislación nacional (Ministerios de Asuntos Económicos y Transformación Digital). Por ello, este estudio contribuirá a la reflexión académica y jurídica sobre cómo regular la IA de manera efectiva, tomando en cuenta las mejores prácticas internacionales (Brynjolfsson & McAfee, 2022).

Asimismo, al comparar los enfoques de Europa, Estados Unidos y China, se podrá evaluar qué modelo es más adecuado para garantizar una IA ética y responsable en el contexto español. Este análisis permitirá valorar el liderazgo de la Unión Europea en la regulación de la IA y su impacto en el resto del mundo (Mozur, 2019).

2. FUNDAMENTOS DE LA INTELIGENCIA ARTIFICIAL

La IA es una disciplina de la informática que permite la creación de sistemas capaces de realizar tareas que tradicionalmente requerían inteligencia humana, tales como el reconocimiento de patrones, la toma de decisiones, la resolución de problemas y el aprendizaje autónomo. Se basa en el desarrollo de algoritmos avanzados, modelos matemáticos y el procesamiento de grandes volúmenes de datos. En la actualidad, la IA es una tecnología clave en diversos sectores, generando importantes oportunidades, pero también nuevos desafíos éticos y jurídicos (Comisión Europea, 2024).

2.1. Definición y evolución de la IA

La IA puede definirse como la capacidad de las máquinas para imitar o replicar funciones cognitivas humanas con el objetivo de mejorar la eficiencia y la precisión en la toma de decisiones (Brynjolfsson & McAfee, 2022).

Existen varias categorías de IA, entre ellas:

- IA Débil o Específica: Diseñada para realizar tareas concretas, como el reconocimiento de voz, la clasificación de imágenes y la traducción automática.

- IA Fuerte o General: Teóricamente capaz de realizar cualquier tarea cognitiva humana, con habilidades de razonamiento, creatividad y toma de decisiones autónoma.
- IA Superinteligente: Una hipotética forma de IA que superaría la inteligencia humana en todos los aspectos y podría mejorar su propio diseño sin intervención humana.

La historia de la IA ha estado marcada por avances tecnológicos y periodos de estancamiento, conocidos como "inviernos de la IA". Su desarrollo puede dividirse en las siguientes etapas:

Años 1950-1960:

- Alan Turing propone la prueba de Turing para determinar si una máquina puede mostrar un comportamiento inteligente (Turing, 1950).
- John McCarthy introduce el término "Inteligencia Artificial" y desarrolla el lenguaje de programación LISP (McCarthy, 1956).
- Creación de los primeros sistemas de resolución de problemas basados en reglas lógicas (Russell & Norvig, 2021).

Años 1970-1980:

- Aparición de los sistemas expertos, programas diseñados para replicar la toma de decisiones de especialistas en campos como la medicina y el derecho (Feigenbaum, 1984).
- Primer "invierno de la IA" debido a la falta de avances en el rendimiento de los modelos computacionales (Nilsson, 2010).

Años 1990-2000:

- Avances en el procesamiento del lenguaje natural y en el reconocimiento de imágenes (Jurafsky & Martin, 2008).
- IBM Deep Blue vence al campeón mundial de ajedrez Garry Kasparov, demostrando la capacidad de la IA para competir con humanos en entornos estructurados (Campbell, Hoane & Hsu, 2002).

Años 2000-2010:

- La IA experimentó en esta década avances significativos en el aprendizaje automático, el procesamiento del lenguaje natural y la robótica. Los principales hitos fueron: Avances en Algoritmos y Modelos de Aprendizaje Automático: Auge del Aprendizaje Profundo y Redes Neuronales. Aunque las redes neuronales artificiales existían desde hace décadas, en los 2000 comenzaron a ganar tracción debido al aumento en la capacidad de cómputo. Primeros avances en traducción automática neuronal: Empresas como Google comenzaron a mejorar los modelos de traducción automática basados en reglas y estadísticas. En 2006, se lanzó Google Translate, utilizando inicialmente redes neuronales para mejorar la traducción (Brynjolfsson & McAfee, 2022).
- Aumento de la Potencia de Cómputo y Big Data: Crecimiento de la Computación Paralela con GPUs. A mediados de los 2000, la computación con unidades de procesamiento gráfico (GPU) comenzó a ser utilizada para entrenar redes neuronales más rápido. NVIDIA y otras empresas promovieron el uso de GPU para acelerar el aprendizaje profundo (Brynjolfsson & McAfee, 2022).

  A medida que Internet crecía, grandes volúmenes de datos comenzaron a ser utilizados para entrenar algoritmos de IA. Empresas como Google, Amazon y Facebook aprovechan la IA para mejorar motores de búsqueda, recomendaciones y publicidad personalizada (Azuaje Pirela & Finol González, 2025).

Años 2010-Actualidad:

- Desarrollo del aprendizaje profundo (*deep learning*) y de redes neuronales convolucionales (LeCun, Bengio & Hinton, 2015).
- AlphaGo de DeepMind derrota a los campeones mundiales de Go, un juego de estrategia más complejo que el ajedrez (Silver et al., 2016).
- Avances en el desarrollo de modelos de IA generativa, como GPT-4 y DALL-E, capaces de producir textos, imágenes y código de manera autónoma (Brown et al., 2020).

2.2. Aplicaciones actuales de la IA

La IA ha revolucionado múltiples sectores, transformando la manera en que se llevan a cabo diversas actividades. Desde la medicina hasta la seguridad, la IA ha demostrado su capacidad para optimizar procesos, mejorar la eficiencia y proporcionar soluciones innovadoras a problemas complejos. A continuación, se presentan algunas de las aplicaciones más relevantes en distintas áreas.

IA en la Medicina:

En el ámbito de la medicina, la IA ha permitido avances significativos en el diagnóstico, tratamiento y gestión de enfermedades. Los algoritmos de aprendizaje profundo han mejorado la capacidad de los sistemas computacionales para detectar patologías a partir de imágenes médicas, como en el caso del diagnóstico automatizado de cáncer de piel mediante redes neuronales convolucionales (Esteva et al., 2017). Además, la IA se ha utilizado en el desarrollo de fármacos predictivos, reduciendo el tiempo y los costos asociados con la investigación de nuevos medicamentos (Murphy, 2012). Otra aplicación relevante es la cirugía robótica, donde los sistemas de IA asisten a los cirujanos en procedimientos complejos, aumentando la precisión y reduciendo los riesgos quirúrgicos (Topol, 2019).

IA en el Derecho y la Justicia:

El uso de la IA en el sector legal ha facilitado el análisis y procesamiento de grandes volúmenes de información jurídica, permitiendo el desarrollo de herramientas para la predicción de sentencias basadas en jurisprudencia previa (Surden, 2014). Esto ha sido particularmente útil en la identificación de patrones en decisiones judiciales y la evaluación de probabilidades de éxito en litigios. Asimismo, la IA se ha aplicado en la automatización de revisiones contractuales, agilizando el análisis de documentos legales y la detección de posibles fraudes o cláusulas abusivas (Ashley, 2017). Otra aplicación emergente es el uso de *chatbots* jurídicos, que proporcionan asesoramiento legal básico a ciudadanos y pequeñas empresas, mejorando el acceso a la justicia de manera eficiente y económica (Sourdin, 2018).

IA en la Educación:

En el ámbito educativo, la IA ha permitido la creación de sistemas de tutoría personalizados, que se adaptan a las necesidades individuales de los estudiantes y optimizan el aprendizaje mediante análisis de desempeño y recomendaciones personalizadas. También se ha implementado en la evaluación automática de exámenes, permitiendo calificaciones objetivas y reduciendo la carga de trabajo de los docentes. Además, plataformas de *e-learning* han incorporado IA para mejorar la experiencia de aprendizaje, proporcionando contenido interactivo y adaptativo según el nivel y ritmo de cada estudiante.

IA en la Seguridad y Defensa:

La IA se ha convertido en una herramienta clave en la seguridad y la defensa, facilitando la implementación de sistemas de reconocimiento facial para la identificación biométrica en aeropuertos, fronteras y sistemas de videovigilancia. También ha sido utilizada para el análisis de patrones de ciberataques, lo que permite detectar amenazas informáticas en tiempo real y reforzar la ciberseguridad de empresas y gobiernos. Otra aplicación destacada es el uso de drones autónomos en operativos de vigilancia y seguridad, que pueden patrullar áreas de alto riesgo y proporcionar información en tiempo real a las fuerzas del orden.

En conclusión, la IA ha transformado múltiples sectores al proporcionar soluciones innovadoras que optimizan procesos y mejoran la eficiencia en diversas disciplinas. No obstante, su implementación plantea desafíos éticos y legales que requieren regulación y supervisión para garantizar su uso responsable (Degli Espositi, 2021).

2.3. Desafíos y oportunidades

El crecimiento acelerado de la IA ha generado numerosos debates sobre sus implicaciones en la sociedad. A continuación, detallamos algunos de los principales desafíos y oportunidades:

Desafíos:

a. Sesgo algorítmico: los sistemas de IA pueden reflejar y amplificar prejuicios existentes en los datos de entrenamiento (Degli Espositi, 2021).
a. Privacidad y protección de datos: la recopilación masiva de información plantea riesgos para la seguridad de los usuarios (Azuaje Pirela & Finol González, 2025).
b. Explicabilidad y transparencia: algunos modelos de IA funcionan como "cajas negras", dificultando la interpretación de sus decisiones (Floridi, 2021).
c. Impacto en el empleo: la automatización de tareas podría provocar desplazamientos laborales en diversos sectores.
d. Uso malintencionado: la IA puede emplearse en actividades delictivas, como fraudes financieros y desinformación masiva (Cortina, 2024).

Oportunidades:

a. Innovación en la salud: la IA permite diagnósticos más rápidos y tratamientos personalizados.
a. Optimización de procesos: mejora la eficiencia en sectores como la industria, la educación y la justicia (Bryson, 2021).
b. Accesibilidad y democratización de la información: herramientas como asistentes virtuales facilitan el acceso a conocimiento y servicios (Azuaje Pirela & Finol González, 2025).
c. Reducción del impacto ambiental: la IA se utiliza para optimizar el consumo de energía y gestionar recursos naturales.

Por lo tanto, vemos como el desarrollo de la IA plantea un equilibrio entre sus beneficios y los riesgos éticos y jurídicos que conlleva. Es fundamental establecer regulaciones adecuadas para garantizar su uso responsable y alineado con los principios democráticos (Cortina, 2024).

3. ÉTICA EN LA INTELIGENCIA ARTIFICIAL

Antes de adentrarnos en el análisis de la ética en el campo de la IA, nos aproximaremos en primer lugar al concepto de ética. ¿Qué es la ética? También llamada filosofía moral, la ética es la disciplina que estudia la conducta humana. Las discusiones éticas se dan en torno al bien y el mal morales, lo correcto y lo incorrecto, la virtud y la idea de deber. Según la RAE la ética es el conjunto de principios y normas que rigen la conducta humana, relacionados con el sentido del bien y del mal (RAE, n.d.). Mientras la moral es el conjunto de principios, juicios o pautas que regulan la conducta humana, la ética es la disciplina que estudia y reflexiona sobre estos mismos preceptos.

El estudio de la ética se remonta a los orígenes mismos de la filosofía en la Antigua Grecia y su desarrollo histórico ha sido amplio y variado. Filósofos como Sócrates (470 a.C. - 399 a.C.), Platón (427 a.C. - 347 a.C.), Aristóteles (384 a.C. - 322 a.C.), San Agustín de Hipona (354 - 430), David Hume (1711 - 1776), Immanuel Kant (1724 - 1804), Arthur Schopenhauer (1788 - 1860) o Friedrich Nietzsche (1844 - 1900) han profundizado en el concepto de la ética. Más recientemente y en España destacaríamos a Julián Marías Aguilera (1914 - 2005), Gustavo Bueno Martínez (1924 - 2016), Eugenio Trías (1942 - 2013), Javier Sábada (1940) o Fernando Savater (1947).

En términos generales, la ética aplicada a la IA que se aborda y propone en el presente trabajo sigue la tradición filosófica ética que va desde la moral individual y social hasta la necesidad de regular y dirigir el desarrollo de la IA en armonía con los principios de justicia, responsabilidad, equidad y demás principios democráticos. Por eso es fundamental que la IA sea diseñada no solo con un objetivo técnico, sino también con una consideración profunda de sus impactos éticos, buscando evitar sesgos, promover la transparencia y garantizar que los derechos humanos no se vean comprometidos.

El concepto de ética utilizado en este estudio no responde a un enfoque único, sino que comprende un marco dinámico integrando distintas corrientes filosóficas que han influido profundamente en el pensamiento occidental actual. Integrando los principios éticos fundamentales para garantizar un uso responsable y justo de la tecnología. Se

busca que la IA sea una herramienta que, no solo beneficie a la humanidad desde un punto de vista técnico y económico, sino que también se desarrolle con responsabilidad y de acuerdo con los principios que se expondrán más adelante, promoviendo una sociedad más equitativa y justa.

Por eso, propondríamos una IA orientada a la búsqueda de la sabiduría, la justicia y la virtud, como sostuvo Sócrates, o encaminada al conocimiento, la bondad y la felicidad como propugnaba Platón. También una IA que trate de defender el bien como último fin, perseverando en tendencias naturales innatas hacia la armonía, la coherencia y el equilibrio como defendió Aristóteles. Que la IA se apoye o se inspire en principios éticos que busquen la verdad y el entendimiento del ser humano como nos enseñó el teólogo latino San Agustín de Hipona.

Son muchos los autores y expertos que, en los últimos tiempos, han abordado esta interesante materia. Luciano Floridi, uno de los mayores expertos en filosofía de la información, afirma que tenemos no solo la oportunidad, sino la obligación de darle a esta nueva herramienta una forma positiva que beneficie a la humanidad y a nuestro planeta.

A continuación, se analizaremos los principales conceptos y principios éticos aplicables a la IA, así como los dilemas más comunes que plantea su uso.

3.1. Concepto de ética en la tecnología

La ética en la tecnología es una rama de la ética aplicada que estudia las implicaciones morales del desarrollo y uso de las tecnologías emergentes. Busca garantizar que las innovaciones tecnológicas respeten valores fundamentales como la justicia, la autonomía, la beneficencia y la no maleficencia (Floridi & Cowls, 2019). A medida que la IA se integra en la toma de decisiones automatizadas, surgen preocupaciones sobre la discriminación algorítmica, la privacidad de los datos y la posibilidad de que los sistemas de IA sean utilizados con fines indebidos o dañinos (Binns, 2018). La creciente

automatización de procesos críticos en sectores como la justicia, la salud y la seguridad subraya la necesidad de establecer marcos éticos sólidos que guíen su diseño e implementación (Mittelstadt et al., 2016).

3.2. Principios éticos aplicables a la IA

Los principios éticos que deben regir la IA han sido objeto de debate en organismos internacionales y académicos. Entre los principios más destacados se encuentran:

a. Transparencia: los sistemas de IA deben ser comprensibles y explicables para los usuarios, garantizando que las decisiones tomadas por los algoritmos puedan ser auditadas y justificadas.

b. Equidad y no discriminación: la IA debe ser diseñada para evitar sesgos que puedan generar discriminación injusta contra determinados grupos sociales, étnicos o de género.

c. Responsabilidad y rendición de cuentas: se deben definir claramente los mecanismos de responsabilidad en el uso de IA, asegurando que exista supervisión humana en la toma de decisiones automatizadas (Gustavo Bueno Martínez, 1996)).

d. Privacidad y protección de datos: la IA debe respetar el derecho a la privacidad y cumplir con regulaciones como el Reglamento General de Protección de Datos (RGPD) en Europa (Parlamento Europeo, 2016).

e. Beneficencia y seguridad: los sistemas de IA deben diseñarse con el objetivo de maximizar los beneficios sociales y minimizar los riesgos potenciales asociados con su uso.

Estos principios han sido adoptados en diversas iniciativas globales, como la Recomendación sobre la Ética de la Inteligencia Artificial de la UNESCO (2021) y las directrices sobre IA confiable de la Comisión Europea (2021).

3.3. Dilemas éticos comunes en sistemas de IA

El desarrollo y uso de la IA han planteado numerosos dilemas éticos que aún no tienen soluciones definitivas. A continuación, se abordan algunos de los principales desafíos éticos que enfrentan los sistemas de IA en la actualidad.

3.3.1. Sesgo algorítmico y discriminación

La inteligencia artificial, lejos de ser un sistema neutral y objetivo, puede reflejar los prejuicios de la sociedad en la que se desarrolla. Aunque muchos consideran que los algoritmos son imparciales, la realidad es que estos modelos aprenden de datos históricos que, en muchos casos, están plagados de desigualdades. Como resultado, las decisiones automatizadas pueden perpetuar y amplificar discriminaciones de género, raza y clase social, con efectos negativos en ámbitos como la contratación laboral, la seguridad, la justicia y el acceso a recursos financieros (Brynjolfsson & McAfee, 2022).

Causas del sesgo algorítmico:

El sesgo en los sistemas de IA puede originarse de múltiples maneras:

1. Datos de entrenamiento desbalanceados: si los datos utilizados para entrenar un modelo no representan de manera equitativa a todos los grupos sociales, los resultados pueden ser sesgados. Por ejemplo, estudios han revelado que los sistemas de reconocimiento facial tienen una tasa de error mucho mayor en personas de piel oscura en comparación con personas de piel clara, debido a conjuntos de datos predominantemente compuestos por imágenes de individuos caucásicos (Buolamwini & Gebru, 2018).
2. Decisiones históricas sesgadas: los algoritmos de IA suelen entrenarse con datos históricos, lo que significa que pueden perpetuar patrones discriminatorios previos (O'Neil, 2016).
3. Diseño de los modelos: algunas decisiones de programación pueden introducir sesgos involuntarios (Bryson, 2021).

Según plantea la investigadora y matemática estadounidense Cathy O'Neil, los algoritmos tienen el peligro de convertirse en "armas de destrucción matemática" cuando no se diseñan con un enfoque ético y de equidad. Esta autora sostiene en su libro Armas de destrucción matemática (*Weapons of Math Destruction*), publicado en 2016, que los modelos de IA están programados para detectar patrones en datos del pasado, lo que significa que, si en el pasado existía discriminación, los sistemas la seguirán reproduciendo. Es decir, si un sistema de selección de personal se basa en datos históricos de contrataciones en el sector tecnológico -dominado por hombres-, es probable que continúe favoreciendo a candidatos masculinos sobre mujeres, sin que haya una intención explícita de discriminar (O'Neil, 2016).

Joy Buolamwini, una informática y activista digital ghanesa-estadounidense, experta en ética de la IA y fundadora de la Algorithmic Justice League, (organización que busca desafiar el sesgo del software en la toma de decisiones), ha evidenciado cómo los sistemas de reconocimiento facial presentan tasas de error significativamente mayores en mujeres y personas racializadas. En su investigación, encontró que la tecnología de grandes empresas tecnológicas como IBM y Microsoft era casi perfecta identificando a hombres blancos, pero fallaba en uno de cada tres casos al analizar rostros de mujeres negras. Esto no solo demuestra que la IA es discriminatoria, sino que además puede generar problemas graves en la vida cotidiana, como identificaciones erróneas en sistemas de vigilancia y control policial (Buolamwini & Gebru, 2018).

Según un informe sobre IA y brechas de género de la UNESCO, el diseño de asistentes virtuales como Alexa, Siri o *Google Assistant* refuerza estereotipos de género problemáticos. Estos sistemas, programados con voces femeninas y respuestas sumisas, pueden consolidar la idea de que las mujeres están destinadas a roles de asistencia y obediencia que supondrían modelos que se están integrando cada vez más en la vida diaria, influyendo en la percepción que las nuevas generaciones pueden tener sobre los roles de género (UNESCO, 2021).

Casos emblemáticos de discriminación algorítmica:

Dos de los casos más controvertidos a nivel mundial por discriminación algorítmica fueron los de Amazon y Compas.

Amazon: uno de los casos más polémicos de sesgo en la IA fue el del sistema de contratación implementado por Amazon, basado en inteligencia artificial. En 2014 la compañía de comercio electrónico puso en marcha un sistema algorítmico con el que esperaba optimizar recursos y ahorrar tiempo y mano de obra, además de encontrar un sistema neutral para contratar personal. El modelo debía evaluar currículums y seleccionar a los mejores candidatos de manera automatizada. Sin embargo, el sistema comenzó a penalizar a las mujeres simplemente porque, en la industria tecnológica, la mayoría de los contratados históricamente habían sido hombres. El algoritmo aprendió de estos patrones y comenzó a descartar solicitudes de mujeres, perpetuando una discriminación histórica sin que nadie lo programara directamente para ello (Dastin, 2018).

Compas: otro caso significativo de sesgo fue el ya mencionado sistema de predicción de reincidencia criminal COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*). Fue una herramienta de gestión de casos y apoyo a la toma de decisiones utilizada por los tribunales de EE. UU. para evaluar la probabilidad de que un acusado se convirtiera en reincidente. Este modelo de IA tenía el doble de probabilidades de clasificar erróneamente a una persona negra como propensa a reincidir en un delito, en comparación con una persona blanca. Lo más preocupante es que jueces y fiscales confiaban en estas evaluaciones para tomar decisiones sobre libertad condicional y sentencias, lo que significaba que los sesgos algorítmicos estaban influyendo directamente en el destino de miles de personas (Angwin, Larson, Mattu, & Kirchner, 2016).

También reseñaremos que en el ámbito financiero, los sistemas de IA utilizados por bancos y empresas de crédito o financieras pueden asignar tasas de interés más altas a personas de ciertos grupos raciales o de bajos ingresos, aunque tengan perfiles financieros similares a los de solicitantes privilegiados. Hace unos años, el método de

pago Apple Card, gestionado por Goldman Sachs atravesó una crisis reputacional tras una denuncia o acusación de que el sistema utilizaba un algoritmo sexista que favorecía a los hombres (Dastin, 2018). Pese a tener una situación crediticia similar, la tarjeta otorga más crédito a los varones que a sus esposas. Se disparó una polémica en redes y se inició una investigación para evaluar las prácticas de este instrumento financiero.

En definitiva, el sesgo algorítmico en la IA puede perpetuar desigualdades y afectar negativamente a grupos vulnerables, dificultando su acceso a oportunidades laborales, justicia equitativa y servicios esenciales (O'Neil, 2016).

Estrategias para mitigar el sesgo algorítmico:

Para reducir el impacto del sesgo en la IA, se han desarrollado estrategias de mitigación;

1. Mejorar la calidad y diversidad de los datos de entrenamiento: es fundamental que los conjuntos de datos utilizados para entrenar modelos de IA sean representativos de toda la población. Esto significa incluir datos de personas de diferentes géneros, edades, razas y contextos socioeconómicos, en lugar de basarse en datos sesgados del pasado.
1. Auditorías algorítmicas constantes: así como las empresas están obligadas a cumplir con regulaciones de privacidad y seguridad, deberían someter sus modelos de IA a auditorías externas para detectar posibles sesgos antes de implementarlos en procesos críticos (Feldstein, 2019).
2. Supervisión humana y explicabilidad: La IA no debe operar como una "caja negra" en la que nadie puede entender cómo se toman las decisiones. Según el científico David Gunning de DARPA (*Defense Advanced Research Projects Agency* o la Agencia de Proyectos de Investigación Avanzados de Defensa, del Departamento de Defensa de los Estados Unidos, responsable del desarrollo de nuevas tecnologías para uso militar), los modelos de IA deben ser diseñados con sistemas de explicabilidad, en los que profundizaremos más adelante, que permitan a los usuarios comprender por qué se llegó a determinada conclusión. Esto es especialmente relevante en ámbitos como la justicia y la salud, donde decisiones erróneas pueden tener consecuencias graves (Gunning & Aha, 2019).

3. Diversidad en los equipos de desarrollo: Un problema recurrente en la industria tecnológica es la falta de mujeres y minorías en los equipos de programación y diseño de IA. Si solo un grupo homogéneo de personas desarrolla estas tecnologías, es más probable que los sistemas reflejen sus propios sesgos. La inclusión de diferentes perspectivas puede ayudar a diseñar modelos más justos y representativos.
4. Regulación y ética en el desarrollo de IA: La Unión Europea ha propuesto el Reglamento de Inteligencia Artificial, que busca establecer controles estrictos para evitar que los sistemas de IA generen discriminación o vulneren derechos fundamentales. Iniciativas como esta son clave para garantizar que la IA beneficie a toda la sociedad, y no solo a quienes tienen el poder de diseñarla y controlarla (Comisión Europea, 2024).

En definitiva, el sesgo en la IA no es un problema técnico menor, sino un reflejo de desigualdades estructurales que deben ser abordadas de manera urgente. La IA tiene el potencial de hacer la sociedad más justa y eficiente, pero si no se corrigen sus sesgos, puede convertirse en una herramienta de discriminación masiva. Por ello, es fundamental implementar regulaciones, auditorías y mecanismos de transparencia para garantizar que estas tecnologías sean justas, equitativas y representativas de toda la sociedad.

3.3.2. Falta de transparencia y explicabilidad

A medida que la IA se convierte en una herramienta clave para la toma de decisiones en ámbitos críticos, como la justicia, la medicina, las finanzas y los recursos humanos, surge una preocupación fundamental: la falta de transparencia en su funcionamiento. Muchos sistemas de IA operan como "cajas negras", es decir, su funcionamiento interno es difícil de interpretar incluso para los propios desarrolladores (Lipton, 2018).

El concepto de explicabilidad se refiere a la capacidad de una IA para justificar sus decisiones de manera comprensible para los humanos. En términos prácticos, significa que los sistemas deben ser diseñados no solo para ofrecer resultados precisos, sino también para hacer transparentes los factores que influyeron en sus decisiones. Esto es

especialmente relevante en modelos basados en aprendizaje profundo (*deep learning*), que procesan información a través de múltiples capas de redes neuronales, generando predicciones altamente precisas, pero difíciles de interpretar (Gunning & Aha, 2019).

Uno de los principales riesgos de la falta de explicabilidad en la IA es su potencial para perpetuar sesgos y discriminaciones sin posibilidad de revisión humana. El caso del ya mencionado software estadounidense COMPAS, fue un claro ejemplo de cómo un sistema opaco puede consolidar desigualdades preexistentes. La falta de explicabilidad del sistema impidió que los acusados pudieran cuestionar o entender las razones detrás de su clasificación, lo que evidencia el peligro de confiar ciegamente en modelos de IA sin mecanismos adecuados de control y auditoría.

La falta de explicabilidad también tiene consecuencias en el ámbito financiero y laboral. Algunos bancos han implementado modelos de IA para evaluar solicitudes de crédito y en el sector tecnológico también se han implementado sistemas de contratación basados en IA como el antes mencionado desarrollado por Amazon que comenzó a penalizar automáticamente los currículums de mujeres. En ambos casos, los sesgos no fueron introducidos intencionalmente, sino que emergieron de datos históricos que reflejaban desigualdades preexistentes. La clave para evitar estos errores no solo radica en mejorar los datos de entrenamiento, sino también en garantizar que las decisiones automatizadas sean comprensibles.

Para abordar estos problemas, la investigación en IA explicable, XAI (startup de IA fundada por Elon Musk) ha desarrollado estrategias para hacer que los modelos sean más interpretables. Una de las soluciones más sencillas es el uso de algoritmos más transparentes, como árboles de decisión o modelos lineales, en lugar de redes neuronales profundas. Otra opción es diseñar sistemas que resalten qué variables fueron determinantes en una decisión. Por ejemplo, si un solicitante de crédito es rechazado, la IA podría especificar que el motivo fue su nivel de ingresos o su historial de pagos, en lugar de simplemente emitir un rechazo sin justificación. También se están desarrollando sistemas que generan explicaciones en lenguaje natural para que los usuarios puedan entender mejor cómo se toman las decisiones.

A nivel regulatorio, el Reglamento General de Protección de Datos de la Unión Europea, en el que más adelante profundizaremos, establece el derecho a obtener explicaciones sobre decisiones automatizadas que afecten a los ciudadanos (Parlamento Europeo, 2021).

Sin embargo, incluso con los avances en explicabilidad, sigue siendo fundamental garantizar la supervisión humana en el uso de la IA. Según el filósofo italiano Luciano Floridi, conocido por sus trabajos sobre la filosofía de la información y la ética informacional, la IA no puede operar de manera autónoma en sectores críticos como la medicina, la justicia o los recursos humanos sin una revisión por parte de especialistas que evalúen sus recomendaciones antes de aplicarlas (Floridi, 2021).

En el ámbito médico, por ejemplo, el cardiólogo y especialista en IA Eric Topol sostiene que, aunque los sistemas inteligentes pueden analizar imágenes médicas y detectar indicios de enfermedades con gran precisión, la decisión final sobre un diagnóstico o tratamiento debe ser tomada por un profesional de la salud, quien puede considerar otros factores clínicos y contextuales que la IA no puede interpretar completamente (Topol, 2019). De la misma manera, en el ámbito judicial, los jueces y abogados deben contar con la capacidad de cuestionar o rechazar las recomendaciones de un algoritmo cuando este influya en la determinación de sentencias, garantizando así que las decisiones sean justas y no estén sesgadas por limitaciones del modelo automatizado.

La supervisión humana en la IA puede darse en distintos niveles. El modelo human-in-the-loop (HITL), traducido al castellano como “humano en el bucle”, hace referencia a sistemas en los que las personas participan activamente en algún paso de cualquier tipo de proceso, generalmente en su supervisión, en la toma de decisiones, o en la valoración y ajuste de sus resultados, asegurando que estos funcionen de manera eficiente, precisa y ética. Por lo tanto, la persona revisa y aprueba la recomendación del sistema antes de tomar una decisión definitiva (Bryson, 2021). En otros modelos, la IA toma decisiones autónomas, pero con la posibilidad de una revisión posterior si se detectan errores o patrones anómalos: human-on-the-loop. Por último, el escenario de mayores riesgos es

aquel en el que la IA opera sin supervisión alguna, human-out-of-the-loop, lo que puede ser aceptable en tareas automatizadas de bajo impacto, pero extremadamente problemático en decisiones que afectan derechos fundamentales.

La tendencia a automatizar cada vez más procesos sin mecanismos claros de explicabilidad y supervisión humana es preocupante. Es innegable que la IA tiene el potencial de mejorar la eficiencia y precisión en muchos sectores, pero delegar decisiones críticas en sistemas que no podemos comprender ni auditar es un riesgo enorme. No podemos permitir que la tecnología se convierta en una herramienta que perpetúe desigualdades de forma silenciosa y sin posibilidad de rendición de cuentas. La verdadera inteligencia no radica solo en desarrollar modelos avanzados, sino en asegurarnos de que estos sean comprensibles, justos y, sobre todo, controlados por los humanos (Brynjolfsson & McAfee, 2022). La IA debe ser una aliada de la sociedad, no una entidad autónoma que tome decisiones sin supervisión ni justificación.

3.3.3. Privacidad y vigilancia masiva

El avance de la IA ha facilitado el desarrollo de tecnologías de vigilancia, como el reconocimiento facial o el análisis de datos de redes sociales. Si bien estas herramientas pueden mejorar la seguridad pública, también plantean riesgos significativos para la privacidad y las libertades civiles.

China, por ejemplo, ha implementado sistemas de vigilancia basados en IA que permiten el monitoreo masivo de su población, lo que ha generado preocupación sobre el uso indebido de estos datos para el control social (Mozur, 2019). En democracias occidentales, el uso de IA en la vigilancia policial ha sido objeto de críticas, ya que estudios han demostrado que estos sistemas podían reforzar sesgos raciales en la identificación de sospechosos.

En respuesta a estas preocupaciones, varias ciudades en Estados Unidos han prohibido el uso de reconocimiento facial en espacios públicos, argumentando que la tecnología

no cumple con estándares adecuados de precisión y protección de derechos civiles (Fussell, 2020).

3.3.4. Desplazamiento laboral y automatización

El avance de la IA y la automatización han generado preocupación sobre el impacto en el empleo. Aunque la IA puede mejorar la productividad y generar nuevas oportunidades, también tiene el potencial de desplazar a trabajadores en múltiples sectores.

Sectores más afectados por la automatización:

De acuerdo con un informe del World Economic Forum (2020), se estimaba que para 2025 la automatización reemplazaría 85 millones de empleos, aunque también podría generar 97 millones de nuevos puestos de trabajo en el ámbito digital. No obstante, el problema radica en la transición laboral y en cómo las economías pueden adaptarse a estos cambios (Brynjolfsson & McAfee, 2022).

Los sectores con mayor riesgo de automatización incluyen:

- Industria manufacturera: robots industriales han reemplazado tareas repetitivas en fábricas, reduciendo la demanda de trabajadores en líneas de ensamblaje.
- Transporte: con el desarrollo de vehículos autónomos, los empleos en el sector del transporte, como conductores de camiones y taxis, podrían verse afectados (Acemoglu & Restrepo, 2019).
- Atención al cliente: los *chatbots* y asistentes virtuales están sustituyendo a trabajadores en centros de llamadas y atención al cliente.

El problema de la desigualdad en la automatización

Uno de los desafíos más importantes es que los trabajadores con menor nivel educativo y formación técnica son los más vulnerables a la automatización. Un informe del McKinsey Global Institute (2021) indicó que el 60% de los empleos que podrían

desaparecer pertenecen a sectores de baja cualificación, mientras que los nuevos empleos creados por la IA requieren habilidades digitales avanzadas.

Este fenómeno podría exacerbar la desigualdad económica, beneficiando a las empresas tecnológicas y a trabajadores altamente capacitados, mientras que otros sectores quedan rezagados (Acemoglu & Restrepo, 2019).

Estrategias para abordar el impacto de la IA en el empleo

Ante este escenario, los expertos han propuesto diversas estrategias para mitigar los efectos negativos de la automatización;

1. Reentrenamiento y formación continua: es fundamental invertir en capacitación digital para que los trabajadores adquieran habilidades tecnológicas. Programas de formación en IA, análisis de datos y programación pueden facilitar la transición laboral.
2. Regulación laboral: algunos países han comenzado a regular el impacto de la IA en el empleo, estableciendo requisitos para la protección de trabajadores desplazados por la automatización (Ministerio de Asuntos Económicos y Transformación Digital, 2021).
3. Fomento de empleos en sectores de alta demanda: la IA está creando nuevas oportunidades en áreas como ciberseguridad, inteligencia de datos y desarrollo de software.

3.3.5. Uso de IA en la guerra y sistemas autónomos letales

El desarrollo de armas autónomas basadas en IA plantea cuestiones éticas y jurídicas sobre la capacidad de las máquinas para tomar decisiones de vida o muerte sin intervención humana. Países como Estados Unidos, China y Rusia están invirtiendo en el desarrollo de sistemas de defensa autónomos, lo que ha generado llamados internacionales para prohibir su uso.

El principal problema radica en la falta de responsabilidad moral de estas tecnologías. Si un dron autónomo ataca un objetivo erróneo, ¿quién es responsable? ¿El programador,

el fabricante o el gobierno que lo desplegó? Estas preguntas han llevado a la ONU a debatir la necesidad de un tratado internacional para regular el uso de armas autónomas (UNESCO, 2021).

Uso de la Inteligencia Artificial en el ámbito militar: sistemas autónomos letales y cuestiones jurídicas

La aplicación de la IA en el ámbito militar constituye uno de los desarrollos más controvertidos y desafiantes de esta tecnología. Entre sus manifestaciones más importantes se encuentran los sistemas de armas autónomos letales (conocidos como *Lethal Autonomous Weapons Systems* o LAWS), capaces de identificar, seleccionar y atacar objetivos sin necesidad de intervención humana directa (Bryson, 2021).

a) Naturaleza y funcionamiento de los sistemas autónomos letales

Estos sistemas combinan sensores, algoritmos de procesamiento de datos y capacidades de ataque, que permiten que una máquina actúe con grados variables de autonomía.

La motivación principal tras el desarrollo de este tipo de armamento reside en el deseo de aumentar la eficacia militar, reducir riesgos para el personal propio y acelerar la capacidad de reacción. Sin embargo, la delegación de decisiones letales a sistemas no humanos cuestiona profundamente la compatibilidad de estas tecnologías con el Derecho Internacional Humanitario (DIH) y el Derecho Internacional de los Derechos Humanos (DIDH) (Floridi, 2021).

b) Principales desafíos jurídicos

El uso de sistemas de armas autónomos letales plantea una serie de cuestiones jurídicas fundamentales;

1. Responsabilidad jurídica: una de las principales cuestiones que se suscitan es la atribución de responsabilidad ante violaciones del DIH o daños a civiles (Russell & Norvig, 2021). La intervención de múltiples actores en el diseño, programación, despliegue y uso de estos sistemas dificulta la identificación clara de un sujeto responsable.

2. Principios del Derecho Internacional Humanitario: el DIH exige el cumplimiento de principios como la distinción entre combatientes y civiles, la proporcionalidad del ataque y la necesidad militar. La capacidad de un sistema automatizado para aplicar estos principios con el mismo grado de juicio y discernimiento que un ser humano es, al menos actualmente, altamente cuestionable.

3. Prohibición de sufrimientos innecesarios: los tratados internacionales prohíben el empleo de armas que causen sufrimientos superfluos o daños desproporcionados. Resulta problemático garantizar que un sistema autónomo pueda evitar este tipo de consecuencias en contextos complejos e impredecibles (Stop Killer Robots, 2020).

4. Derechos humanos y control humano significativo: desde la óptica del DIH, la intervención humana directa en decisiones letales es esencial para garantizar el derecho a la vida y el principio de dignidad humana (ACAMI, 2024). En este contexto, emerge el concepto de "*meaningful human control*", que aboga por mantener un control humano efectivo sobre las decisiones críticas (Artificial Intelligence, Human Rights, Democracy, and the Rule of Law, 2021).

c) Estado actual de la regulación internacional

Hasta el momento, no existe un tratado internacional específico que regule o prohíba expresamente los sistemas autónomos letales. Pero se han promovido iniciativas dentro del marco de la Convención sobre Ciertas Armas Convencionales (CCW) de las Naciones Unidas, que ha servido de foro para el debate sobre su posible regulación.

Algunos Estados y organizaciones han solicitado una prohibición preventiva de estos sistemas, mientras que otros, particularmente aquellos con mayor capacidad tecnológica y militar, se inclinan por enfoques más flexibles y basados en la autorregulación.

d) Consideraciones éticas y jurídicas de futuro

El desarrollo y uso de sistemas autónomos letales plantea un riesgo considerable de deshumanización del conflicto armado y de erosión de la responsabilidad jurídica. La comunidad internacional se enfrenta al reto de garantizar que el progreso tecnológico no transgreda los límites establecidos por el derecho internacional y los principios fundamentales del humanitarismo (Floridi, 2021).

El caso de la guerra en Ucrania, iniciada en febrero de 2022 y aún activa, ha sido un escenario clave para el uso de la IA en el ámbito militar, con aplicaciones de drones autónomos para ataques, ciberdefensa, y análisis masivo de datos. Si bien estas tecnologías han mejorado la eficiencia y proporcionado ventajas tácticas, también han planteado las serias preocupaciones éticas y jurídicas. Las implicaciones jurídicas incluyen la dificultad para atribuir responsabilidad en caso de violaciones de derechos humanos, el cumplimiento del Derecho Internacional Humanitario y la manipulación informativa mediante IA, como los *deepfakes*. La ausencia de un marco normativo internacional adecuado subraya la extrema urgencia de establecer regulaciones claras que aseguren el uso ético de la IA en conflictos bélicos.

Este conflicto ha mostrado tanto el potencial como los riesgos de la IA en la guerra, destacando la necesidad de una supervisión humana adecuada para evitar que la tecnología amplifique injusticias. Es fundamental que el desarrollo de la IA esté respaldado por regulaciones estrictas y principios éticos para garantizar que sirva a la humanidad de manera justa y responsable.

Los dilemas éticos en la IA evidencian la necesidad de una regulación más estricta y un enfoque ético en el diseño y aplicación de estas tecnologías. La transparencia, la equidad y la responsabilidad deben ser principios rectores en la evolución de la IA. Si bien la IA tiene el potencial de mejorar la sociedad, su desarrollo debe estar alineado con los valores democráticos y derechos humanos.

4. REGULACIÓN JURÍDICA DE LA IA EN LA UNIÓN EUROPEA

La regulación de la IA en la Unión Europea ha sido una prioridad en los últimos años, debido al impacto que esta tecnología tiene en los derechos fundamentales, la economía y la seguridad. La Comisión Europea ha adoptado un enfoque basado en la protección de los valores democráticos, buscando garantizar un desarrollo responsable de la IA que combine la innovación con la salvaguarda de los derechos de los ciudadanos (Comisión Europea, 2024).

El marco jurídico europeo se ha construido a partir de documentos clave, como el Libro Blanco sobre la Inteligencia Artificial (2020) y el Reglamento Europeo de Inteligencia Artificial (AIA, por sus siglas en inglés)**,** aprobado en 2023. Este reglamento se convierte en el primer marco legal integral que regula la IA a nivel global, estableciendo normas estrictas para su desarrollo y aplicación en la Unión Europea. A lo largo de este apartado, analizaremos sus disposiciones, su evaluación desde una perspectiva ética y su impacto en los Estados miembros.

4.1. Análisis del Reglamento de Inteligencia Artificial

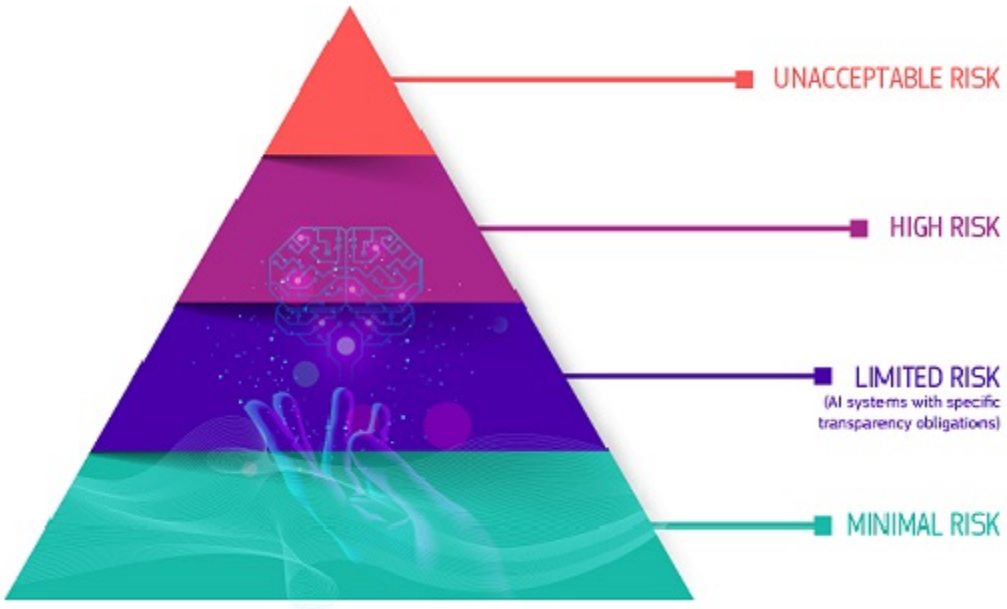
El Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n° 300/2008, (UE) n° 167/2013, (UE) n° 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial) que se analizará a continuación se publicó en el Diario Oficial de la Unión Europea (DOUEL) el 12 de Julio de 2024 y entró en vigor el 1 de agosto de 2024. Constituye el primer marco jurídico integral para la IA en la Unión Europea y su propósito es garantizar el desarrollo y uso de la IA de manera ética, segura y alineada con los valores fundamentales de la UE, como la protección de los derechos humanos, la seguridad y la transparencia (Faggiani & Garrido Carrillo, 2025)**.**

La normativa responde a la necesidad de evitar la fragmentación regulatoria entre los Estados miembros, estableciendo normas armonizadas para el desarrollo, comercialización y uso de sistemas de IA en la Unión. Se enfoca en prevenir riesgos y fomentar la innovación responsable, garantizando al mismo tiempo la libre circulación de bienes y servicios basados en IA en el mercado interior. Para ello, introduce categorías de riesgo, prohíbe ciertos usos problemáticos y regula estrictamente aplicaciones de alto impacto.

Principales aspectos del Reglamento de Inteligencia Artificial

El Reglamento Europeo de Inteligencia Artificial introduce una de las clasificaciones más detalladas y estructuradas hasta la fecha para evaluar el impacto y los riesgos de los sistemas de IA en la sociedad. A diferencia de otros marcos regulatorios más generales, la UE ha optado por un enfoque basado en el nivel de riesgo, permitiendo que la regulación se aplique de manera diferenciada en función de las posibles consecuencias de cada sistema. Esto garantiza que se impongan restricciones a las aplicaciones que pueden afectar los derechos fundamentales, mientras que se permite una mayor flexibilidad en aquellas que tienen un impacto limitado.

A continuación, se detallan las cuatro categorías o niveles de riesgo establecidos por el Reglamento:



A. Sistemas de IA de riesgo inaceptable

Los sistemas de riesgo inaceptable son aquellos que presentan un peligro claro y grave para los derechos fundamentales y, por lo tanto, han sido prohibidos en la Unión Europea. La UE ha establecido una lista de aplicaciones de IA que entran en esta categoría, entre ellas;

- Sistemas de manipulación subliminal que pueden influir en el comportamiento de las personas sin su conocimiento o consentimiento. Esto incluye tecnologías diseñadas para modificar la voluntad de un individuo de manera encubierta, lo que podría ser utilizado para la manipulación política o comercial sin que el usuario sea consciente de ello.
- Sistemas de puntuación social, similares al modelo implementado en China, donde el comportamiento de los ciudadanos es monitorizado y evaluado para determinar su acceso a servicios públicos o privados. Este tipo de IA plantea serios riesgos para la privacidad, la autonomía personal y la discriminación social, razones por las cuales la UE ha optado por su prohibición total.
- Sistemas de vigilancia masiva basados en reconocimiento facial, salvo en circunstancias excepcionales. La UE ha sido muy estricta en la regulación del uso de biometría en espacios públicos, ya que este tipo de tecnología podría ser utilizada para controlar el comportamiento ciudadano o limitar la libertad de expresión y movimiento. Sin embargo, se han establecido ciertas excepciones para casos de seguridad pública y prevención de delitos graves, aunque su uso estará sujeto a una estricta supervisión (Feldstein, 2019).

La prohibición de estas aplicaciones representa un posicionamiento firme de la UE en la defensa de los derechos fundamentales y la protección de la privacidad, diferenciándola de otras potencias globales que han permitido la proliferación de estas tecnologías sin restricciones claras (Feldstein, 2019).

B. Sistemas de IA de alto riesgo

Los sistemas de IA considerados de alto riesgo son aquellos que pueden afectar significativamente la seguridad, la salud o los derechos fundamentales de las personas.

Esta categoría incluye aplicaciones utilizadas en sectores críticos donde una decisión errónea o sesgada de la IA podría generar consecuencias negativas graves.

Los ámbitos clave donde se aplica esta categoría son;

- Sanidad: sistemas de IA empleados en diagnóstico médico, cirugía asistida por robots o evaluación de historiales clínicos. La regulación exige que estos modelos sean auditados antes de su uso, garantizando que no introduzcan sesgos que afecten la calidad de la atención médica. Además, se establece que siempre debe haber una supervisión humana que valide las decisiones tomadas por la IA (Topol, 2019).
- Recursos humanos: herramientas de IA utilizadas en procesos de selección de personal, evaluaciones de desempeño y despidos. Para evitar sesgos discriminatorios, las empresas estarán obligadas a demostrar que sus sistemas cumplen con criterios de transparencia y equidad (Bryson, 2021).
- Control migratorio y fronterizo: aplicaciones utilizadas en la detección de mentiras, evaluación de riesgo en fronteras o identificación biométrica. Dada la sensibilidad de estos sistemas, la UE exige que se realicen pruebas rigurosas antes de su implementación.
- Justicia y seguridad: sistemas que analizan la probabilidad de reincidencia criminal, toman decisiones judiciales automatizadas o asisten en investigaciones policiales. La regulación exige que estos sistemas sean explicables y que sus decisiones puedan ser revisadas por un ser humano.

Los desarrolladores de IA de alto riesgo deberán cumplir con requisitos específicos, como la documentación del proceso de desarrollo, la implementación de mecanismos de supervisión humana, la auditoría independiente y la transparencia en la toma de decisiones. Además, estarán sujetos a controles regulares para garantizar que no generen discriminación o efectos adversos no previstos (Comisión Europea, 2024).

C. Sistemas de IA de riesgo limitado

En esta categoría se incluyen los sistemas de IA que interactúan directamente con los usuarios, pero cuyo impacto en los derechos fundamentales es reducido. No se exige una regulación tan estricta como en los casos de alto riesgo, pero sí se han impuesto ciertas obligaciones de transparencia para garantizar que los ciudadanos sean conscientes de que están interactuando con una IA (Comisión Europea, 2024).

Algunos ejemplos de aplicaciones en esta categoría incluyen;

- *Chatbots* y asistentes virtuales como Alexa, Siri o *Google Assistant*.
- Sistemas de recomendación utilizados en plataformas de *streaming*, comercio electrónico y redes sociales.
- Filtros de contenido y moderación en plataformas digitales, como los algoritmos de TikTok o YouTube que deciden qué contenido mostrar a los usuarios.

Las principales obligaciones que deben cumplir estos sistemas incluyen la notificación clara de que se está interactuando con una IA y la posibilidad de intervención humana en ciertas circunstancias. Por ejemplo, si un usuario de una plataforma de atención al cliente desea ser atendido por una persona en lugar de un *chatbot*, la empresa debe garantizar esa opción.

D. Sistemas de IA de riesgo mínimo

Esta categoría engloba las aplicaciones de IA que no suponen un riesgo significativo para la sociedad y que, por lo tanto, no están sujetas a regulaciones estrictas. Se trata de herramientas que operan en segundo plano y cuyo impacto es meramente funcional.

Ejemplos de IA de riesgo mínimo incluyen;

- Filtros de spam en correos electrónicos.
- Sistemas de IA utilizados en videojuegos, como la generación de NPCs (*non playing caracters* o personajes no jugadores) inteligentes.

- Automatización de procesos simples, como sistemas de IA utilizados en hojas de cálculo para organizar datos.

Dado que estos sistemas no afectan directamente los derechos de las personas, no están sujetos a auditorías ni a supervisión específica (Bryson, 2021).

La clasificación de la IA en función del nivel de riesgo es una de las estrategias más acertadas del Reglamento Europeo de IA, ya que permite una regulación diferenciada en función del impacto real de cada tecnología. Mientras que se han impuesto prohibiciones absolutas a las IA que pueden ser peligrosas para los derechos fundamentales, como la puntuación social y la manipulación subliminal, también se han mantenido normas flexibles para aquellas aplicaciones cuyo impacto es limitado o nulo.

Esta clasificación es equilibrada y necesaria, ya que permite proteger a los ciudadanos sin frenar la innovación en sectores donde la IA puede aportar valor sin generar riesgos significativos. Sin embargo, el éxito de esta normativa dependerá en gran medida de la capacidad de supervisión de la UE y los Estados miembros, asegurando que las normas realmente se cumplan y no queden en simples declaraciones de intenciones. Además, será clave encontrar un equilibrio entre regulación y competitividad para que las empresas europeas puedan seguir innovando sin verse asfixiadas por una carga burocrática excesiva (Faggiani & Garrido Carrillo, 2025).

Obligaciones y sanciones

El Reglamento impone estrictas obligaciones a los desarrolladores y usuarios de IA, especialmente en las aplicaciones de alto riesgo. Entre ellas se incluyen la necesidad de;

- Garantizar explicabilidad y transparencia en los sistemas de IA.
- Implementar mecanismos de supervisión humana para evitar decisiones automáticas que afecten derechos fundamentales (Topol, 2019).
- Llevar a cabo evaluaciones de impacto antes de lanzar al mercado una IA de alto riesgo (Bryson, 2021).

- Registrar y documentar adecuadamente los sistemas de IA para su auditoría y control.

El Reglamento no solo establece un marco normativo para el desarrollo y uso de la IA, también introduce un régimen de sanciones diseñado para garantizar su cumplimiento. La Unión Europea ha optado por un sistema de multas proporcionales al nivel de gravedad de la infracción (Comisión Europea, 2024).

El objetivo de este régimen sancionador es disuadir a las empresas y administraciones públicas de implementar sistemas de IA sin los controles adecuados, protegiendo así los derechos fundamentales de los ciudadanos europeos. Además, se busca evitar ventajas competitivas y desleales para aquellas empresas que decidan ignorar la regulación en favor de una innovación sin restricciones (Bryson, 2021).

Mecanismos de supervisión y aplicación de sanciones

Para garantizar el cumplimiento del Reglamento, la UE ha establecido un sistema de supervisión descentralizado, en el que cada Estado miembro será responsable de aplicar la normativa dentro de su territorio. Cada país deberá designar una autoridad nacional de supervisión de la IA, encargada de realizar auditorías, investigar denuncias y aplicar sanciones cuando sea necesario.

A nivel europeo, se ha creado la Oficina Europea de Inteligencia Artificial. Ésta apoya el desarrollo y el uso de una IA fiable, al tiempo que protege contra los riesgos de la IA. La Oficina se creó como centro de conocimientos especializados en IA y constituye la base de un sistema europeo único de gobernanza de la IA. El Reglamento o la llamada Ley de IA es el primer marco jurídico global sobre IA en todo el mundo, que garantiza la salud, la seguridad y los derechos fundamentales de las personas y proporciona seguridad jurídica a las empresas de los 27 Estados miembros (Faggiani & Garrido Carrillo, 2025).

Impacto de las sanciones en las empresas y el mercado tecnológico

Las elevadas sanciones establecidas por el Reglamento de IA han generado preocupación en el sector tecnológico. Por un lado, algunas grandes tecnológicas han criticado que estas multas pudieran desincentivar la innovación en IA dentro de Europa, provocando que muchas compañías trasladen su desarrollo a regiones con regulaciones más laxas, como Estados Unidos o Asia. Existe el riesgo de que, en lugar de fomentar un entorno competitivo, la UE termine limitando la capacidad de las empresas europeas para competir con gigantes tecnológicos como Google, Amazon o Alibaba. Pero por otro lado, las multas también pueden servir como una barrera contra prácticas abusivas, evitando que empresas lancen productos de IA sin los controles adecuados. Esto es especialmente importante en áreas como la contratación laboral, la justicia y los servicios financieros, donde un uso irresponsable de la IA podría reforzar desigualdades y discriminaciones preexistentes.

Desde una perspectiva jurídica, el hecho de que las sanciones sean proporcionales al nivel de gravedad de la infracción es un punto a favor del Reglamento. No todas las violaciones de la normativa deben ser castigadas con el mismo rigor, y esta diferenciación permite que las empresas tengan la oportunidad de corregir errores menores sin enfrentarse a multas desproporcionadas (Faggiani & Garrido Carrillo, 2025).

Por todo esto, el régimen sancionador del Reglamento Europeo de Inteligencia Artificial es uno de los más estrictos a nivel mundial, reflejando el compromiso de la UE con la protección de los derechos fundamentales y la responsabilidad en el desarrollo de IA. Sin embargo, su éxito dependerá de cómo se implementen las sanciones y de si realmente se logra un equilibrio entre control y fomento de la innovación. Las sanciones elevadas son necesarias para evitas abusos pero deben aplicarse de manera justa y con mecanismos de supervisionen efectivos, si la normativa se convierte en una carga burocrática excesiva, podría frenar la competitividad de las empresas europeas frente a otros mercados. La clave estará en hacer cumplir la regulación sin desincentivar la inversión y el desarrollo tecnológico en Europa.

4.2. Evaluación ética del marco normativo europeo

El Reglamento Europeo de Inteligencia Artificial ha sido diseñado con el propósito de garantizar que el desarrollo y uso de la IA se alineen con los valores fundamentales de la Unión Europea, como la protección de los derechos humanos, la seguridad, la transparencia y la equidad (Comisión Europea, 2024).

Desde un punto de vista ético, este marco regulador representa un esfuerzo significativo para equilibrar el progreso tecnológico con la responsabilidad social, estableciendo límites claros para evitar abusos y garantizar que la IA opere dentro de un marco de derechos fundamentales. Uno de los principales aciertos del Reglamento es su énfasis en la transparencia y la explicabilidad, especialmente en los sistemas de IA de alto riesgo. Esta falta de explicabilidad ha sido objeto de críticas debido a que puede perjudicar a ciudadanos que no tienen forma de impugnar decisiones automatizadas que les afectan directamente (Brynjolfsson & McAfee, 2022).

Otro aspecto clave del Reglamento es su prohibición del uso de IA para vigilancia masiva y manipulación de la opinión pública. Esta medida busca prevenir la implantación de sistemas de reconocimiento facial en espacios públicos, salvo en casos excepcionales relacionados con la seguridad nacional y la prevención del terrorismo. La Unión Europea ha sido muy clara en su postura contra modelos de control social basados en IA, como el sistema de crédito social implementado en China, donde los ciudadanos son puntuados en función de su comportamiento y pueden ver restringidos sus derechos si obtienen una calificación baja. Este tipo de tecnologías plantean graves riesgos éticos, ya que pueden utilizarse para limitar la libertad de expresión, la privacidad y la autonomía personal.

Algunos sectores han expresado su preocupación de que las restricciones impuestas a la IA en Europa sean excesivas y puedan frenar la competitividad y la innovación en el sector tecnológico. A pesar de estas críticas, la Comisión Europea sostiene que una IA ética y transparente puede convertirse en un valor diferencial que fortalezca la confianza del consumidor y evite problemas futuros derivados de un uso irresponsable de la

tecnología. Si bien la regulación impone ciertos costos adicionales para las empresas, a largo plazo podría generar un mercado más seguro y sostenible, donde la IA sea confiable y beneficiosa para la sociedad (Floridi & Cowls, 2019).

4.3. Impacto en los Estados miembros

El Reglamento Europeo de Inteligencia Artificial tiene un impacto significativo en los Estados miembros, ya que introduce una legislación uniforme para toda la UE y establece reglas claras sobre cómo debe ser el desarrollo, la comercialización y el uso de la IA en el territorio comunitario. Esta armonización de normas es fundamental para evitar la fragmentación regulatoria, es decir, que cada país desarrolle su propia legislación sobre IA de manera independiente, lo que podría generar barreras en el mercado digital y dificultar la cooperación transnacional.

Uno de los principales desafíos para los Estados miembros es la creación de organismos nacionales de supervisión encargados de garantizar el cumplimiento del reglamento. Cada país deberá establecer autoridades competentes en IA, que se encargarán de auditar, fiscalizar y sancionar el uso indebido de la IA dentro de su territorio. Estas entidades trabajarán en coordinación con la Oficina Europea de Inteligencia Artificial, cuyo rol de supervisión a nivel comunitario asegurará que la normativa se aplique de manera coherente en todos los países de la UE.

En el caso de España, el gobierno ha tomado la delantera en la implementación del Reglamento con la creación de la Agencia Española de Supervisión de la Inteligencia Artificial (AESIA), convirtiéndose en uno de los primeros países en establecer un organismo específico para la supervisión de la IA. La estrategia española busca fomentar una IA ética y segura, alineada con los valores democráticos de la UE, pero sin perder de vista la necesidad de impulsar la competitividad del sector tecnológico (Ministerio de Asuntos Económicos y Transformación Digital, 2021).

Otros países, como Francia y Alemania, han manifestado su interés en fortalecer la inversión en IA para mantener la competitividad europea frente a Estados Unidos y China. Ambos países han abogado por un enfoque equilibrado que combine regulación y flexibilidad, permitiendo que las empresas tecnológicas tengan margen para innovar sin que la burocracia frene su crecimiento. Sin embargo, algunos Estados miembros con economías más pequeñas han expresado su preocupación sobre los costes de implementación de la regulación, ya que cumplir con las nuevas normativas podría suponer un reto financiero considerable para startups y empresas emergentes.

En términos generales, el impacto del Reglamento dependerá de su implementación efectiva y de la capacidad de la UE para garantizar un entorno equilibrado entre innovación y regulación (Faggiani & Garrido Carrillo, 2025). Es esencial que los Estados miembros trabajen en conjunto y armonizados para que la IA en Europa pueda desarrollarse dentro de un marco seguro, ético y competitivo.

La armonización normativa que impone el Reglamento es un paso importante para garantizar que la IA en la UE se desarrolle con estándares éticos elevados. Cada país enfrentará retos particulares en su aplicación, y será crucial encontrar un equilibrio entre control y crecimiento. Algunos sectores critican que esta regulación pudiera ralentizar la innovación, es innegable que marca un precedente global en la gobernanza de la IA, la clave estará en implementar medidas que protejan los derechos fundamentales sin generar un marco regulatorio tan restrictivo que frene la competitividad y la innovación tecnológica en Europa (Cortina, 2024).

La UE corre el riesgo de convertirse en líder en regulación, pero no en desarrollo de IA, y eso podría hacer que las empresas tecnológicas europeas queden rezagadas frente a las de otras regiones o países, pero como ya expusimos antes, un punto a valorar altamente del Reglamento es la clasificación de los sistemas de IA en función de su nivel de riesgo. No tiene sentido regular con la misma intensidad un filtro de *spam* que un algoritmo de reconocimiento facial usado por la policía. Al establecer categorías de riesgo, se permite una regulación flexible que no frena completamente la innovación en aplicaciones de bajo impacto. Esto demuestra que la UE no está tratando de limitar el

desarrollo de la IA, sino de asegurarse de que se use de forma ética y controlada cuando su impacto en la sociedad es significativo (Azuaje Pirela & Finol González, 2025).

Otro aspecto que genera dudas es la capacidad real de los Estados miembros para hacer cumplir el Reglamento. No basta con aprobar leyes ambiciosas si luego no existen los recursos y las estructuras adecuadas para supervisar su cumplimiento. Se han propuesto agencias nacionales de supervisión de IA, lo cual es positivo, pero habrá que ver si realmente tienen los medios para auditar modelos complejos, detectar violaciones y sancionar a quienes no cumplan con las normas. Si la implementación de esta regulación no es eficiente, podría convertirse en una normativa con buenas intenciones, pero de difícil aplicación. El Reglamento es un avance importante porque protege a los ciudadanos de los riesgos asociados a la IA, pero será clave observar cómo evoluciona su aplicación en los próximos años, asegurándose de que las restricciones no se conviertan en un obstáculo para el progreso tecnológico.

5. ESTRATEGIA ESPAÑOLA EN MATERIA DE IA

5.1. Objetivos y pilares de la Estrategia Nacional de IA

El Consejo de Ministros aprobó el pasado 14 de mayo de 2024, a propuesta del Ministerio para la Transformación Digital y de la Función Pública, la Estrategia de Inteligencia Artificial 2024 (Ministerio para la Transformación Digital y de la Función Pública, 2024). Esta estrategia da continuidad a las iniciativas desplegadas por el Gobierno de España hasta el momento en materia de IA, adaptándolas a los notables cambios experimentados en esta tecnología en los últimos años (Gobierno de España, 2024).

Según se informa en la página oficial del Ministerio de Economía, Comercio y Empresa se trata de un plan ambicioso, diseñado para consolidar y expandir el uso de la IA en el conjunto de la economía y en la administración pública y su despliegue se realizará entre los años 2024 y 2025. Contará con recursos por importe de 1.500 millones de

euros, adicionales a los 600 millones ya movilizados, procedentes fundamentalmente del Plan de Recuperación, Transformación y Resiliencia y de la adenda a dicho Plan.

La Estrategia de Inteligencia Artificial 2024 está estructurada en tres grandes ejes que activarán ocho palancas de acción.

Eje 1: Refuerzo de las capacidades para el desarrollo de la IA

Este eje de la estrategia pone el acento en la necesidad de reforzar las palancas para el desarrollo de la IA para ser capaces de aprovechar lo máximo posible las oportunidades que ofrece esta tecnología, centrándose en cuatro elementos: el refuerzo de la supercomputación, la capacidad de almacenamiento sostenible, los modelos de lenguaje y la necesidad de talento.

En primer lugar, en lo relativo a la supercomputación, contempla la inversión de 90 millones de euros para la puesta en marcha de nuevos clústeres altamente especializados, que permitan mejorar las prestaciones del MareNostrum 5, del Barcelona SupercomputingCenter - Centro Nacional de Supercomputación (BSC-CNS) y el refuerzo de la Red Española de Supercomputación. En segundo lugar, la estrategia incluye iniciativas para el establecimiento de Centros de Procesamiento de Datos (CPD) ambientalmente sostenibles, a través de un nuevo marco regulatorio que mejore la planificación de estas infraestructuras y simplifique los trámites administrativos, además de ordenar territorialmente la implantación de futuros CPD.

En tercer lugar, este eje incluye la creación y expansión de una familia de modelos de lenguaje en castellano y lenguas cooficiales que se llamará ALIA que permitirán reducir los sesgos y mejorar las aplicaciones prácticas que las empresas y administraciones de nuestro país pueden desarrollar. Finalmente, la cuarta palanca de este eje es el impulso del talento especializado en IA en un contexto en el que existe una gran demanda de profesionales (Gobierno de España, 2024).

Eje 2: Facilitar la aplicación de la IA en el sector público y privado

En el marco de este eje se articulará un procedimiento para la implantación de la IA en la Administración General de Estado (AGE), a través del proyecto GobTech Lab, que canalizará los casos de uso de esta tecnología a través de un laboratorio de innovación para desarrollar proyectos piloto y soluciones innovadoras para las entidades del sector público estatal.

En segundo lugar, para promover el desarrollo de la IA en el sector privado, especialmente en el ámbito de las pymes y autónomos, se desarrollará el programa Kit Consulting, dotado con 300 millones de euros, para que los proyectos empresariales de menor tamaño puedan contratar servicios de asesoramiento para la adopción de la IA. Para las iniciativas contempladas en los ejes 1 y 2 (palanca 4 y 6), se invertirá hasta 300 millones procedentes de los Fondos FEDER (Gobierno de España, 2024).

Eje 3: Fomentar una IA transparente, ética y humanística

El cumplimiento de los objetivos de este eje se articula a través de la Agencia Española de Supervisión de la Inteligencia Artificial (AESIA), que actua en una triple dirección: como centro de pensamiento y análisis sobre la IA (analizando tendencias, generando debate social e identificando buenas prácticas y riesgos emergentes), como supervisor de un despliegue responsable de la IA (certificando sistemas de IA de acuerdo con el reglamento europeo de IA y estableciendo buenas prácticas para promover modelos transparentes y abiertos) y como referente internacional (participando en las instituciones europeas y mundiales de gobernanza de la IA) (Gobierno de España, 2024).

5.2. Anteproyecto de Ley para el buen uso y la gobernanza de la Inteligencia Artificial e integración de principios éticos en la estrategia española

Hace unos días, el pasado 11 de marzo de 2025, el Consejo de Ministros ha aprobado éste Anteproyecto que adaptará la legislación española al Reglamento Europeo de IA,

ya en vigor, combinando un enfoque regulador con el impulso a la innovación. El Anteproyecto busca garantizar un uso de la IA ético, inclusivo y beneficioso para las personas (Gobierno de España, 2025).

Se tramitará por la vía de urgencia y seguirá los trámites preceptivos antes de volver al Consejo de Ministros para su aprobación definitiva como Proyecto de Ley y posterior envío a las Cortes Generales para su aprobación.

Con el objetivo de que la Unión Europea disponga de un marco legal común para el desarrollo, comercialización y uso de sistemas de IA que eviten los riesgos para las personas, se prohíben determinados usos maliciosos de la IA, se introducen obligaciones más rigurosas para sistemas considerados de alto riesgo y se establecen unos requisitos mínimos de transparencia para el resto.

Incorpora, además, un nuevo derecho digital de retirada provisional del mercado español de sistemas de IA por la autoridad de vigilancia competente cuando hayan provocado un incidente grave, como el fallecimiento de una persona (El Confidencial, 2025).

Prácticas prohibidas de la IA

Las prácticas prohibidas entraron en vigor el 2 de febrero de 2025 y desde el 2 de agosto de 2025 se podrán sancionar mediante multas u otras medidas adicionales (requerir su adaptación al sistema para que sea conforme, impedir que se comercialice…) aplicando el régimen sancionador que incorpora el Anteproyecto de Ley, dentro de las horquillas que fija el reglamento europeo. A modo de ejemplo, son prácticas prohibidas;

- El uso de técnicas subliminales (imágenes o sonidos imperceptibles) para manipular decisiones sin consentimiento, causando un perjuicio considerable a la persona (adicciones, violencia de género o menoscabo de su autonomía). (Ej: un *chatbot* que identifica usuarios con adicción al juego y les incita a entrar, con técnicas subliminales, en una plataforma de juego online).

- Explotar vulnerabilidades relacionadas con la edad, la discapacidad o situación socioeconómica para alterar sustancialmente comportamientos de modo que les provoque o pueda provocar perjuicios considerables (Ej: un juguete infantil habilitado con IA que anima a los niños a completar retos que les producen o pueden producirles daños físicos graves).
- La clasificación biométrica de las personas por raza u orientación política, religiosa o sexual. (Ej: un sistema de categorización facial biométrica capaz de deducir la orientación política o sexual de un individuo mediante análisis de sus fotos en redes sociales).
- La puntuación de individuos o grupos basándose en comportamientos sociales o rasgos personales como método de selección para, por ejemplo, denegarles la concesión de subvenciones o préstamos.
- Valorar el riesgo de que una persona cometa un delito basándose en datos personales como su historial familiar, nivel educativo o lugar de residencia, con excepciones legales.
- Inferir emociones en centros de trabajo o educativos como método de evaluación para promoción o despido laboral, salvo por razones médicas o de seguridad.

Las autoridades encargadas de vigilar los sistemas prohibidos serán la Agencia Española de Protección de Datos (para sistemas biométricos y gestión de fronteras); el Consejo General del Poder Judicial (para sistemas de IA en el ámbito de la justicia), la Junta Electoral Central (para sistemas que IA que afecten a procesos democráticos) y la Agencia Española de Supervisión de la Inteligencia Artificial en el resto de los casos (El Confidencial, 2025).

Recientemente, se ha celebrado el I Encuentro de Derecho Digital del Ilustre Colegio de Abogados de Madrid (ICAM) en el que se han presentado las directrices de la corporación madrileña en un completo plan estratégico para abordar el uso de la IA en la profesión jurídica, destacando la participación activa del ICAM en el proceso regulatorio actualmente en desarrollo en España. Los cuatro ejes vertebradores formación especializada para el uso eficiente de la IA, participación en el desarrollo normativo, promoción de estándares éticos y creación de un centro de asesoría en IA

para despachos y abogados. Además, el Colegio publicará una guía con las directrices para garantizar la supervisión humana, la transparencia y la protección de derechos en el uso de esta tecnología en el sector legal (El Diario de Madrid, 2025).

5.3. Comparación con otras estrategias nacionales

El enfoque de la estrategia española en materia de IA alineado con los principios de Reglamento Europeo sobre Inteligencia Artificial busca garantizar la protección de los derechos fundamentales, la transparencia y la seguridad en su uso, implementando un marco ético robusto para enfrentar los numerosos riesgos de esta tecnología. Pero en otros países se adoptan enfoques particulares, adaptados a sus contextos políticos, económicos y sociales, que afectan significativamente al desarrollo y regulación de la IA. En el siguiente punto se mencionarán las estrategias de Estados Unidos y China, las dos potencias pioneras en esta materia, pero, ¿qué hay de los otros países?

Reino Unido adoptó su estrategia nacional de IA en 2021, priorizando la transparencia y la responsabilidad en el desarrollo de la IA. Su enfoque enfatiza el fomento de la innovación mientras que también promueve una IA ética y centrada en el ser humano. En particular, se destaca la importancia de prevenir los sesgos y garantizar que la IA no refuerce desigualdades preexistentes. Esta estrategia se complementa con la creación de un Consejo de Ética de la IA para supervisar el impacto social de la tecnología y asegurar que las aplicaciones de IA sean justas y transparentes (Gobierno de Reino Unido, 2021).

En Canadá, el Instituto Canadiense de Investigación en Inteligencia Artificial (CIFAR) establece directrices que promueven una IA inclusiva y responsable. La estrategia canadiense busca equilibrar la innovación tecnológica con la protección de los derechos humanos y la privacidad. En 2021, el gobierno canadiense presentó su Estrategia Nacional de IA, que destaca principios como la responsabilidad social, la equidad y la transparencia. Canadá también ha sido pionero en iniciativas para garantizar que las auditorías de IA sean accesibles y transparentes, permitiendo que los

ciudadanos comprendan las decisiones tomadas por los sistemas automatizados (CIFAR, 2023).

México, a pesar de estar en una etapa menos avanzada en el desarrollo de su estrategia de IA, ha comenzado a implementar políticas que buscan fomentar la innovación tecnológica y al mismo tiempo proteger los derechos fundamentales. El Consejo Nacional de Ciencia y Tecnología (CONACYT), en colaboración con la Secretaría de Economía, ha propuesto el desarrollo de directrices éticas para la adopción de IA en el sector público y privado, con énfasis en la transparencia, la protección de la privacidad y la inclusión social. Aunque aún no existe una ley nacional que regule completamente la IA, se están dando pequeños pasos hacia la creación de un marco normativo que regule su uso de forma responsable (CONACYT, 2023).

Brasil ha adoptado una estrategia nacional de IA que integra principios éticos centrados en la inclusión social y el desarrollo sostenible. La estrategia brasileña busca garantizar que la IA sea utilizada para reducir desigualdades sociales, con especial enfoque en la educación y la salud pública. En 2020, el gobierno brasileño aprobó la *Estrategia Brasileira de Inteligência Artificial* (EBIA), que establece directrices para la gestión ética de la IA, destacando la importancia de la privacidad, la transparencia y la no discriminación. Además, Brasil está trabajando en la creación de un marco normativo para regular el uso de IA en el sector público y privado, buscando un equilibrio entre innovación y responsabilidad social (Gobierno de Brasil, 2020).

En India, el gobierno ha comenzado a abordar la ética de la IA en el contexto de su estrategia digital. Aunque la regulación formal aún está en desarrollo, India ha promovido principios de innovación inclusiva y acceso equitativo a la tecnología, buscando evitar la exclusión digital. El país está impulsando políticas para fomentar el uso de la IA en sectores como la agricultura, la salud y la educación, con un énfasis en su uso para resolver problemas sociales y mejorar las condiciones de vida (Gobierno de la India, 2023).

6. PERPECTIVAS INTERNACIONALES SOBRE LA ÉTICA EN LA IA

6.1. Enfoque de Estados Unidos en la ética de la IA

Estados Unidos sigue un enfoque descentralizado y basado en la autorregulación del sector privado, con principios que equilibran la innovación con la protección de los derechos individuales. Los principios clave de la ética de la IA en Estados Unidos son;

1. Énfasis en los derechos individuales y la privacidad: a diferencia de China, EE. UU. prioriza la protección de la privacidad de los ciudadanos. Existen regulaciones como la Ley de Privacidad del Consumidor de California y la Ley de Portabilidad y Responsabilidad del Seguro Médico que protegen el uso de datos en sectores específicos.
2. Regulación descentralizada y basada en el mercado: no existe una ley federal unificada sobre IA, sino regulaciones sectoriales desarrolladas por organismos como la Comisión Federal de Comercio (FTC) y el Departamento de Comercio. Se confía en la autorregulación de las empresas tecnológicas, con directrices voluntarias para el desarrollo ético de la IA (FTC, 2024).
3. Fomento de la innovación y la competitividad global: el gobierno busca evitar regulaciones que frenen la innovación y el liderazgo de EE. UU. en IA frente a China. Se han promovido incentivos para la investigación en IA responsable, con financiación a proyectos que abordan los riesgos éticos (Center for Strategic and International Studies [CSIS], 2023).
4. Seguridad nacional y geopolítica: la IA es vista como un activo estratégico en la competencia con China. Se han impuesto restricciones a la exportación de chips avanzados y modelos de IA a China para proteger la seguridad nacional (White & Case LLP, 2024).
5. Regulación emergente de la IA generativa: la administración Biden dictó el 30 de octubre de 2024, una Orden Ejecutiva para gestionar los riesgos de la IA (*Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence)* que estableció nuevos estándares para el desarrollo de una IA segura, protección de la privacidad de los estadounidenses, promoción de la equidad y los derechos

civiles, defensa de los consumidores y trabajadores, promoción de la innovación y la competencia, promover el liderazgo estadounidense en materia de IA y garantizar el uso gubernamental responsable y eficaz de la IA (Reuters, 2024). Pero el nuevo Presidente de los EE.UU., Donnald Trump, ha firmado en enero de 2025 un nuevo decreto para impulsar sistemas de IA "libres de sesgos ideológicos o agendas sociales diseñadas" y que deroga la según él "peligrosa" normativa establecida por su predecesor, Joe Biden, sobre seguridad y transparencia en el sector. La FTC (Comisión Federal de Comercio de Estados Unidos) está investigando el impacto de la IA en la desinformación y la equidad: la apertura de una investigación recayó sobre cinco de las grandes tecnológicas sobre sus estrategias de inversión y alianzas en el desarrollo de la IA: Alphabet, Amazon, Microsoft, OpenAI y Anthropic, convertidas en pocos años en las empresas con más peso de EE.UU. El primer argumento que ofrece la FTC es que necesita "una mejor comprensión interna de estas relaciones y su impacto en la competencia", para averiguar si las políticas de "estas compañías dominantes ponen en peligro la innovación y socavan la libre competencia" (Center for Strategic and International Studies, 2023).

6.2. Enfoque de China en la ética de la IA

China considera la IA como una herramienta clave para el crecimiento económico, la estabilidad social y la seguridad nacional. Su enfoque ético está definido por un fuerte control estatal y una alineación con los valores del Partido Comunista. Los principios clave de la ética de la IA en China son;

1. Gobernanza centralizada y regulación estricta: el gobierno chino supervisa y regula el desarrollo de la IA a través de organismos como la Administración del Ciberespacio de China (CAC) y el Ministerio de Ciencia y Tecnología**.** Se han establecido normativas que obligan a las empresas a cumplir con estándares de transparencia y seguridad en el desarrollo de algoritmos (Bandurski, 2023).
2. Ética alineada con el socialismo y el bienestar colectivo: la AI debe promover el desarrollo económico y la estabilidad social. Se enfatiza la responsabilidad de las empresas tecnológicas para garantizar que la IA no genere efectos adversos en la

sociedad. Se aplican restricciones a contenidos generados por IA que puedan desafiar la ideología del Estado.

3. Uso de la IA para el control social y la seguridad: china ha implementado sistemas de vigilancia basados en IA, incluyendo reconocimiento facial y monitoreo ciudadano mediante el Sistema de Crédito Social. Las plataformas de redes sociales utilizan IA para censurar contenidos que el gobierno considera problemáticos (China Media Project, 2023).
4. Control sobre datos y privacidad: aunque China promulgó la Ley de Protección de Información Personal (PIPL) en 2021, que otorga ciertos derechos de privacidad a los ciudadanos, el Estado mantiene amplias facultades para acceder a datos por razones de seguridad nacional (China Law Translate, 2017).
5. Regulación estricta de la IA generativa: desde 2023, China ha establecido normativas que exigen que el contenido generado por IA no atente contra la estabilidad social ni infrinja los valores del Partido Comunista. Las empresas deben registrar sus modelos de IA con el gobierno y someterse a auditorías de seguridad.

Principales diferencias entre China y EE. UU. UU.

China y EE.UU. UU. presentan modelos opuestos en la ética de la IA. Mientras que China prioriza la estabilidad social y el control estatal, con un fuerte énfasis en la supervisión gubernamental, EE.UU. favorece la innovación con una regulación descentralizada, enfocándose en la privacidad y la competitividad.

Coincidencias entre China y EE.UU.

Pero a pesar de sus diferencias, ambos países comparten ciertos principios en la ética de la IA;

- Reconocimiento de la IA como una tecnología estratégica: tanto China como EE.UU. ven la IA como un activo clave para el desarrollo económico y la seguridad nacional y han invertido fuertemente en el desarrollo de IA para aplicaciones militares y de defensa (Liu, 2023).

- Regulación de los riesgos: ambos gobiernos han comenzado a desarrollar normativas para prevenir riesgos de la IA generativa, como la manipulación de la información y la discriminación algorítmica. Existen esfuerzos para regular la transparencia y la seguridad de los modelos de IA (Zhuang, 2024).
- Enfoque en la transparencia algorítmica: China y EE.UU. UU. buscan mayor transparencia en los sistemas de IA para garantizar su correcto funcionamiento. En EE.UU., esto se traduce en auditorías y estándares de equidad en algoritmos, mientras que en China implica control gubernamental sobre las decisiones algorítmicas.
- Promoción de la IA responsable: ambos países han publicado documentos con principios éticos para la IA, destacando la necesidad de equidad, seguridad y responsabilidad (Zhuang, 2024).

Por lo tanto, parece claro que el futuro de la ética de la IA dependerá de cómo estos dos líderes tecnológicos equilibren la regulación e innovación, en un contexto de creciente competencia global y preocupaciones por el uso de la IA en la geopolítica y la seguridad.

7. PROPUESTA DE PRINCIPIOS ÉTICOS PARA LA REGULACIÓN JURÍDICA DE LA IA EN ESPAÑA

Analizaremos el documento de la Comisión Europea sobre Directrices Éticas para una IA fiable. Podemos dividir estas directrices en dos categorías; por una lado, los fundamentos de una IA fiable reflejados en 4 principios éticos, y por otro, los 7 requisitos clave para la realización de esta IA fiable, de forma que se garantice el cumplimiento de los requisitos clave.

7.1 Identificación de principios clave

El Capítulo 1 de las Directrices, donde se establecen los cuatro principios éticos fundamentales para fundamentar una IA fiable, explica que éstos son clave para

garantizar que la IA se desarrolle de manera que respete los derechos y la dignidad humana, así como los valores fundamentales de la sociedad.

1. Respeto de la autonomía humana

Según este principio los sistemas de IA deben permitir y fomentar la autonomía de los seres humanos. La IA debe ser diseñada de manera que empodere a las personas para tomar decisiones informadas y para controlar los procesos en los que interactúan con la tecnología. Este respeto a la autonomía humana implica que los individuos no deben ser dominados ni manipulados por sistemas automatizados, sino que deben poder tomar decisiones con pleno conocimiento de causa. La IA debe ser una herramienta que facilite la autonomía personal, ayudando a las personas en la toma de decisiones, pero sin sustituir su capacidad de elegir de manera libre y consciente. El principio de autonomía también está vinculado al derecho a la autodeterminación de los individuos, que deben poder decidir si desean interactuar con sistemas de IA y bajo qué condiciones.

2. Prevención del daño

El segundo principio ético clave en las Directrices se refiere a la prevención del daño causado por los sistemas de IA. Este principio exige que los sistemas de IA sean diseñados para minimizar los riesgos y daños potenciales, tanto a nivel individual como social. Los desarrolladores de IA deben garantizar que los sistemas sean seguros, robustos y capaces de manejar situaciones imprevistas para evitar daños a las personas, al medio ambiente o a las comunidades. La prevención del daño también está relacionada con la idea de la responsabilidad, ya que, si se producen daños derivados de un mal funcionamiento de la IA, debe existir una responsabilidad clara sobre las consecuencias.

3. Equidad

El principio de equidad establece que los sistemas de IA deben ser diseñados y operados de manera que respeten los principios de justicia, igualdad y no discriminación. Los algoritmos de IA no deben generar sesgos ni reforzar prejuicios existentes, y deben

garantizar que todas las personas, independientemente de su género, raza, origen étnico, orientación sexual o discapacidad, sean tratadas de manera equitativa. La equidad en la IA también implica que los sistemas deben ser accesibles y justos para todos, lo que requiere una regulación activa que asegure que no se perpetúen desigualdades. Este principio tiene como objetivo prevenir la discriminación algorítmica, que puede surgir si los sistemas de IA son entrenados con datos sesgados o si no se implementan medidas adecuadas para garantizar la imparcialidad. Asegurar la equidad en la IA contribuye a promover una sociedad más inclusiva y justa.

4. Explicabilidad

La explicabilidad es el principio que exige que los sistemas de IA sean transparentes y que sus decisiones sean comprensibles para los seres humanos. Las personas deben poder entender cómo y por qué se han tomado decisiones que les afectan. Este principio es especialmente importante en aplicaciones de IA que tienen un impacto significativo en la vida de las personas, como la justicia, la atención médica y los servicios financieros. Los usuarios deben tener acceso a explicaciones claras sobre los algoritmos y los modelos que subyacen a las decisiones automatizadas.

Estos cuatro principios éticos garantizan que la IA sea eficiente y útil, y que se alinee con los valores democráticos y humanos, promoviendo la justicia, la equidad y la transparencia.

En España, estos principios ya se han tenido en cuenta en el desarrollo normativo, asegurando que la IA se desarrolle y utilice de acuerdo con valores democráticos y constitucionales.

7.2. Recomendaciones para la incorporación en la legislación española

La incorporación de los principios éticos en la regulación jurídica de la IA en España responde no solo a una necesidad técnica y jurídica, sino también a una reflexión profunda sobre el papel que la tecnología debe jugar en la vida humana y social. Son

siete los principios clave que garantizan una IA fiable, no solo como imperativos legales, sino como valores fundamentales que respetan los derechos de las personas y el bien común.

Acción y supervisión humanas

El principio de acción y supervisión humanas implica una cuestión filosófica esencial: el poder humano sobre la máquina. No puede ser delegado el poder en una inteligencia que carezca de consciencia o valor moral. Esta debe ser la columna vertebral de los sistemas de IA, garantizando que las decisiones fundamentales, especialmente aquellas que afectan a derechos humanos básicos, sigan siendo tomadas por personas. Esto se relaciona con el principio de autonomía (que presupone que las personas son agentes de su propio destino) y con la dignidad humana, que debe ser preservada frente a la lógica fría de un sistema automatizado. Así, la legislación española debe asegurar que cualquier IA de alto impacto permita la intervención de las personas, garantizando la justicia y la no arbitrariedad.

Solidez técnica y seguridad

La solidez técnica y seguridad reflejan la responsabilidad ética que debe recaer sobre los desarrolladores de sistemas de IA. Para garantizar la protección de los derechos fundamentales y de la seguridad jurídica, los sistemas deben ser resilientes, fiables y estar respaldados por planes de contingencia. La legislación española debe exigir auditorías de seguridad rigurosas y protocolos para la reparación de fallos, especialmente en sectores de alto riesgo, como la sanidad y la justicia.

Gestión de la privacidad y de los datos

La privacidad y la protección de los datos entroncan también con otros aspectos de la ética en la IA. La privacidad es un derecho humano fundamental que debe ser garantizada con normas claras y eficaces. La autonomía de los individuos debe ser respetada, permitiendo que éstos tengan control sobre su información personal y sean informados adecuadamente de cómo se utilizará.

Transparencia

Otro de los principios más robustos en materia de IA, como se ha analizado a lo largo de todo el trabajo es la transparencia. La legislación española debería promover el uso de tecnologías de trazabilidad y la obligación de explicar de manera accesible el funcionamiento y las decisiones de los sistemas de IA.

Diversidad, no discriminación y equidad

Los principios de diversidad, no discriminación y equidad subrayan la importancia de crear sistemas de IA que no perpetúen sesgos o injusticias, sino que promuevan la inclusión y el respeto a la dignidad de todos los individuos. La igualdad ante la ley y la prohibición de discriminación en cualquiera de sus formas, ya sea por género, raza, orientación sexual, discapacidad u otras características personales son esenciales.

Bienestar social y ambiental

El principio de bienestar social y ambiental conceta con la responsabilidad ética de desarrollar tecnologías que no solo beneficien a los usuarios individuales, sino que contribuyan positivamente al bienestar colectivo. La legislación española debería fomentar el uso de la IA para abordar problemas globales como la pobreza, la salud pública o la inmigración.

Rendición de cuentas

Finalmente, el principio de rendición de cuentas promueve un marco de responsabilidad clara para los desarrolladores y usuarios de sistemas de IA. La legislación española debería garantizar que todos los sistemas de IA sean auditables, con procesos claros para que las decisiones automatizadas puedan ser cuestionadas y corregidas por un ser humano, con sanciones y mecanismos de rendición de cuentas proporcionales a la gravedad de los daños causados.

8. CONCLUSIONES

A lo largo de este trabajo de Fin de Grado se ha analizado el indiscutible valor que la IA tiene en la sociedad actual pero también los importantes riesgos y desafíos que estos sistemas plantean en el ámbito ético, jurídico y de seguridad. Se ha estudiado la actual regulación y el enfoque más adecuado desde la perspectiva ética así como los límites que, eventualmente, deben operar.

La IA, con su potencial para transformar la vida cotidiana, plantea cuestiones éticas esenciales que deben ser protegidas para garantizar que sus beneficios sean equitativos y respetuosos con los derechos humanos. La rápida expansión de la IA exige una reflexión filosófica profunda sobre cómo debe ser regulada. ¿Qué limites éticos debería marcar la legislación para evitar traspasar fronteras hacia terrenos perjudiciales y dañinos para el hombre? Deben, por supuesto, protegerse los derechos fundamentales y las libertades individuales y es crucial que principios éticos como la autonomía humana, la equidad, la no discriminación, la transparencia y la responsabilidad en el diseño y aplicación de la IA, guien la regulación de esta materia, dando así cumplimiento a los valores democráticos de nuestro ordenamiento jurídico.

Con el Reglamento Europeo de IA, la Unión Europea ha tomado la delantera en la regulación de esta tecnología. No obstante, los desafíos siguen siendo considerables, y su éxito dependerá de la implementación efectiva de las normas y de cómo los Estados miembros logren garantizar que la IA se use de manera ética y responsable.

España se ha alineado con las directrices de la Unión Europea. La estrategia nacional de IA refleja este compromiso, pero la clave será afrontar de manera efectiva la regulación normativa que garantice el cumplimiento de los principios antes mencionados y una continua evaluación de los sistemas de IA, asegurando que no se utilicen para fines perjudiciales o injustos. Por eso, ética y derecho deben caminar de la mano. La legislación de esta materia debe construirse sobre la base de los principios éticos que fundamentan nuestro derecho. Y es que, en esencia, nuestro derecho es ético. Los desarrolladores de IA deben ser responsables desde una perspectiva legal, a la vez que

ética, considerando las consecuencias sociales y humanas de sus creaciones, estando atentos a garantizar un perfecto -aunque a veces complicado- equilibrio entre la innovación y la protección de los derechos.

En un mundo donde la IA juega un papel cada vez más protagonista, su integración en todos los aspectos de la vida social, política y económica es necesaria por lo que debe existir un diálogo constante entre tecnología y ética. El uso de IA ya ha generado situaciones como la automatización del empleo o la ampliación de la brecha digital y conforme vayan apareciendo más problemáticas será fundamental ajustar la normativa para que pueda responder a estos desafíos de manera efectiva.

El estudio de la ética de la IA y su regulación está aún en una etapa incipiente, por lo que es fundamental que continúen los esfuerzos de investigación en áreas como la prevención del sesgo algorítmico, el análisis de las implicaciones sociales de la automatización o la mejora de la explicabilidad de los sistemas de IA. Además, la investigación interdisciplinaria entre filosofía, derecho, ciencia y tecnología es crucial para crear soluciones justas y equilibradas a los problemas éticos de la IA. Con una mirada ética a la inteligencia artificial desde el campo de la filosofía del derecho, aseguraremos que la IA no solo sea avanzada, sino también beneficiosa para el hombre.

9. BIBLIOGRAFÍA

ACAMI. (2024). *Ética y control humano significativo en sistemas de armas autónomos letales regidos por la Inteligencia Artificial*. https://www.acami.es/wp-content/uploads/2024/04/Etica-y-control-humano-significativo-en-SALAS-IA-web.pdf

Acemoglu, D., & Restrepo, P. (2019). Automation and New Tasks: How Technology Displaces and Reinstates Labor. *Journal of Economic Perspectives, 33*(2), 3-30. https://doi.org/10.1257/jep.33.2.3

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. *ProPublica.* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Artificial Intelligence, Human Rights, Democracy, and the Rule of Law. (2021). *Artificial Intelligence, Human Rights, Democracy, and the Rule of Law: A Primer*. https://arxiv.org/abs/2104.04147

Atienza Navarro, M. L. (2022). *Daños causados por inteligencia artificial y responsabilidad civil*. Cuadernos de Derecho de Daños.

Azuaje Pirela, M., & Finol González, D. (2025). *Introducción a la ética y el derecho de la inteligencia artificial*. TEMAS LA LEY.

Bandurski, D. (2023, abril 14). Bringing AI to the Party. *China Media Project*. https://en.wikipedia.org/wiki/Cyberspace_Administration_of_China

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics, 143*(1), 30-56. https://doi.org/10.1016/j.jfineco.2021.05.031

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, 149-159. https://doi.org/10.1145/3287560.3287583

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.

Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/1607.06520

Brynjolfsson, E., & McAfee, A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies.* W. W. Norton & Company.

Brynjolfsson, E., & McAfee, A. (2022). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W. W. Norton & Company.

Bryson, J. (2021). *The Ethics of Artificial Intelligence: Principles, Challenges, and Global Perspectives*. Oxford University Press.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)*. https://doi.org/10.1145/3278721.3278776

Center for Strategic and International Studies. (2023, septiembre 12). *Se avecina una regulación de la IA: ¿Cuál es el resultado probable?* https://www.csis.org/blogs/strategic-technologies-blog/ai-regulation-coming-what-likely-outcome

Comisión Europea. (2019). *Directrices éticas para una inteligencia artificial confiable.* https://ec.europa.eu/digital-strategy/our-policies/european-approach-artificial-intelligence

China Law Translate. (2017). Social Credit System: Not Just Another Chinese Idiosyncrasy. *Princeton Journal of East Asian Studies*. https://jpia.princeton.edu/news/social-credit-system-not-just-another-chinese-idiosyncrasy

China Media Project. (2023, abril 14). Bringing AI to the Party. https://en.wikipedia.org/wiki/Cyberspace_Administration_of_China

Comisión Europea. (2024). *Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52024PC1689

Cortina, A. (2024). *¿Ética o ideología de la inteligencia artificial?: El eclipse de la razón comunicativa en una sociedad tecnologizada*. Ediciones Paidós.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters.* https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Degli Espositi, S. (2021). *La ética de la inteligencia artificial*. Los Libros de la Catarata.

El Confidencial. (2025). *El Gobierno aprueba una ley que obliga a identificar el contenido generado con IA*. El Confidencial. https://www.elconfidencial.com/tecnologia/2025-03-11/gobierno-uso-etico-ia-obliga-identificar-contenido-generado_4083723/

El Derecho. (2020, noviembre 18). Aprobado el anteproyecto de ley para el buen uso y la gobernanza de la inteligencia artificial. *El Derecho*. https://elderecho.com/aprobado-el-anteproyecto-de-ley-para-el-buen-uso-y-la-gobernanza-de-la-inteligencia-artificial

El Diario de Madrid. (2025). *El ICAM impulsa la regulación responsable de la Inteligencia Artificial en el ámbito jurídico*. El Diario de Madrid. https://www.eldiariodemadrid.es/articulo/actualidad/icam-impulsa-regulacion-responsable-inteligencia-artificial-ambito-juridico/20250320165547094536.html

El País. (2019, noviembre 26). Los riesgos y las oportunidades de la inteligencia artificial para las finanzas. *El País*. https://elpais.com/economia/2019/11/26/finanzas_a_las_9/1574784148_993972.html

European Commission. (2021). *Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

European Parliament. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation - GDPR). https://eur-lex.europa.eu/eli/reg/2016/679/oj

Faggiani, V., & Garrido Carrillo, F. J. (Eds.). (2025). *Avances en el desarrollo de la inteligencia artificial: Principios éticos y democráticos, concreciones normativas en la Unión Europea, y límites para las Administraciones públicas* (1st ed.). Colección Estudios.

Feldstein, S. (2019). *The Rise of Digital Repression: How Technology is Reshaping Power, Politics, and Resistance*. Oxford University Press.

Floridi, L. (2021). *Ética de la inteligencia artificial*. Herder Editorial.

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., et al. (2018). *AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations.* Minds and Machines, 28(4), 689-707.

FTC. (2024). *FTC announces crackdown on deceptive AI claims, schemes*. https://www.ftc.gov/news-events/press-releases/2024/09/ftc-announces-crackdown-deceptive-ai-claims-schemes

Fussell, S. (2020). Why Cities Are Banning Facial Recognition Technology. *The Atlantic.* https://www.theatlantic.com/technology/archive/2020/06/why-cities-are-banning-facial-recognition-technology/613075/

Gobierno de España. (2025). *El Gobierno da luz verde al anteproyecto de ley para un uso ético, inclusivo y beneficioso de la Inteligencia Artificial*. Ministerio para la Transformación Digital. https://digital.gob.es/comunicacion/sala-de-prensa/comunicacion_ministro/2025/03/2025-03-11.html

Gunning, D., & Aha, D. W. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine, 40*(2), 44-58. https://doi.org/10.1609/aimag.v40i2.2850

Gustavo Bueno Martínez. *El sentido de la vida. Seis lecturas de filosofía moral* (1996).

IEEE. (2020). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

Kurzweil, R. (2025). *La singularidad está más cerca: Cuando nos fusionamos con la IA*. Deusto.

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue, 16*(3), 31-57.

Liu, Q. (2023, julio 11). China to lay down AI rules with emphasis on content control. *Financial Times*. https://holisticai.com/blog/china-ai-regulation

Ministerio de Asuntos Económicos y Transformación Digital. (2021). *Estrategia Nacional de Inteligencia Artificial ()*. Gobierno de España. https://portal.mineco.gob.es/es-es/ministerio/areas-prioritarias/Paginas/inteligencia-artificial.aspx

Mozur, P. (2019). One Month, 500,000 Face Scans: How China Is Using AI to Profile a Minority. *The New York Times.* https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group.

Parlamento Europeo. (2024). *Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo de 13 de junio de 2024 sobre normas armonizadas para la inteligencia artificial*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52024PC1689

Real Academia Española. (n.d.). *Ética*. En *Diccionario de la lengua española* (23.ª ed.). https://dle.rae.es

Reuters. (2024, septiembre 15). *FTC investigates major tech firms over AI ethics and competition*. https://www.reuters.com/article/tech-ai-competition-ethics-ftc-investigation

Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking Press.

Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*. Pearson.

Stop Killer Robots. (2020). *Los riesgos de las armas autónomas: Una perspectiva ética y legal sobre la inteligencia artificial en la guerra*. https://www.stopkillerrobots.org/wp-content/uploads/2020/09/Los-riesgos-de-las-armas-autonomas-una-perspec-min.pdf

Topol, E. (2019). Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books.

Turing, A. M. (1950). *Computing Machinery and Intelligence*. *Mind*, 59(236), 433–460. https://academic.oup.com/mind/article/LIX/236/433/986238

UNESCO. (2021). *Recomendación sobre la ética de la inteligencia artificial*. Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO). https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa

Villani, C. (2018). *For a Meaningful Artificial Intelligence: Towards a French and European Strategy*. https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf

White & Case LLP. (2024, diciembre 18). *AI Watch: Rastreador global de regulaciones - Estados Unidos*. https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united-states

Zhuang, R. (2024, septiembre 15). China's AI Policy & Development: What You Need to Know. *FiscalNote*. https://fiscalnote.com/blog/china-ai-policy-development-what-you-need-to-know